DUOYUAN TONGJI GAILUN YU SHIYAN

多元统计概论与实验

刘桂梅　林伟然　编著

$$P(i/x)=\frac{q_i f_i(x)}{\sum_{j=1}^{m} q_j f_j(x)},\quad i=1,2,\mathrm{L}\ \ m$$

ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

多元统计概论与实验

DUOYUAN TONGJI GAILUN YU SHIYAN



ISBN 978-7-308-11991-7

定价：25.00元

多元统计概论与实验

刘桂梅　林伟然　编著


ZHEJIANG UNIVERSITY PRESS
浙江大学出版社

图书在版编目（CIP）数据

多元统计概论与实验 / 刘桂梅，林伟然编著．—杭州：浙江大学出版社，2013.8
ISBN 978-7-308-11991-7

Ⅰ．①多… Ⅱ．①刘… ②林 Ⅲ．①多元分析—高等学校—教材 Ⅳ．①O212.4

中国版本图书馆 CIP 数据核字（2013）第 184498 号

多元统计概论与实验

刘桂梅　林伟然　编著

责任编辑　杜希武
封面设计　刘依群
出版发行　浙江大学出版社
（杭州市天目山路 148 号　邮政编码 310007）
（网址：http://www.zjupress.com）
排　　版　杭州好友排版工作室
印　　刷　浙江云广印业有限公司
开　　本　787mm×1092mm　1/16
印　　张　12
字　　数　292
版 印 次　2013 年 8 月第 1 版　2013 年 8 月第 1 次印刷
书　　号　ISBN 978-7-308-11991-7
定　　价　25.00 元

浙江大学出版社发行部联系方式：(0571) 88925591，http://zjdxcbs.tmall.com

前　　言

多元统计分析是经典统计学的一个重要分支，内容十分丰富、应用范围极为广泛。它能够在多个对象和多个指标互相关联的情况下，应用数理统计学分析方法提炼出规律性的结论。特别是，随着电子计算机的普及和软件的发展，信息储存手段以及数据信息的成倍增长，多元分析的方法已广泛应用于自然科学和社会科学的各个领域。国内外实际应用中卓有成效，已证明了多元分析方法是处理多维数据不可缺少的重要工具，并日益显示出无比的魅力。

为了满足应用型学院的有关专业开设“多元统计分析”课程的要求，作者根据多年教学实践经验编写了本书，希望给应用型学院的学生提供一本既能保持多元统计理论的系统性，又能理论联系实际，进行案例数据处理和分析的教科书。在系统介绍多元统计分析基本理论和方法的同时，选择典型的事例进行剖析，注重介绍每种多元分析方法的实际背景、统计思想、统计模型、数学原理和解题的思路，突出实际问题的应用和统计思想的渗透。

本教材分为理论和实验两部分，每章都有数学实验环节，结合实例介绍统计软件的操作和使用，理论性和实用性相结合，提高学生学习兴趣，培养学生解决实际问题的能力。配备适当课后练习和思考题，鼓励学生自己利用一些实际数据进行操作和实现，巩固所学知识。实验的每一个步骤都有贴图演示，以及输出结果的详细分析，帮助学生理解和掌握所学知识，做到“学以致用”。

本书共分十章，主要介绍经典多元分析的基本理论，包括多元正态及其抽样分布、假设检验；聚类分析、判别分析、主成分分析、因子分析、典型相关分析等主要的多元统计方法。理论课与实验课的内容按 2∶1 的模式设计，通过对实际案例的数据分析培养学生统计思想，熟练应用所学的统计方法和各种统计模型，培养学生建立统计模型解决问题的能力，提高学生的综合素质。

本书编写过程中，作者参考了高等院校常用的教材和近年来的有关文献资料。在写作中，作者力求做到语言简洁流畅，层次清楚，理论与实践操作并重，内容丰富，既便于学生循序渐进地系统学习多元统计的基本理论，又能使他们熟悉掌握这些理论的应用和实际数据的处理，解决实际问题。本书不仅可以作为应用型学院有关专业的教材，也可以供广大实际工作者、科研人员阅读和参考。

本书作为应用型学院“多元统计分析”的教材，根据我们多年的教学实践，安排 32＋16 课时，可以讲授前八章。教师可以根据具体情况，灵活选讲。

目 录

第一篇 多元统计分析原理与方法

第一章 绪论 …… 3
1.1 多元统计简介 …… 3
1.2 主要内容安排 …… 3
第二章 多元数据图表示法 …… 5
2.1 散点图矩阵 …… 5
2.2 雷达图 …… 7
2.3 调和曲线图 …… 7
2.4 脸谱图 …… 9
第三章 均值向量和协方差阵的检验 …… 12
3.1 随机向量 …… 12
3.2 多元正态分布 …… 13
3.2.1 多元正态分布的定义及基本性质 …… 13
3.2.2 多元正态分布的参数估计 …… 14
3.3 均值向量的检验 …… 15
3.3.1 单个正态总体 $N_p(\mu,\sum)$ 均值向量的检验 …… 15
3.3.2 两个正态总体 $N_p(\mu_1,\sum_1)$ 和 $N_p(\mu_2,\sum_2)$ 均值向量的检验 …… 16
3.3.3 多个正态总体均值向量的检验(多元方差分析) …… 17
3.4 协差阵的检验 …… 18
3.4.1 一个正态总体协方差阵的检验 …… 18
3.4.2 多个正态总体协方差阵的检验 …… 18
第四章 聚类分析 …… 23
4.1 距离 …… 23
4.1.1 聚类数据的标准化处理 …… 23
4.1.2 样品距离的定义 …… 24

4.2 系统聚类法…… 25
4.2.1 类间的距离…… 25
4.2.2 四种系统聚类法…… 26
4.3 K-均值聚类法 …… 31

第五章 判别分析 …… 33

5.1 判别分析简介…… 33
5.2 距离判别法…… 33
5.2.1 两组距离判别…… 34
5.2.2 多个总体的距离判别法…… 36
5.3 贝叶斯(Bayes)判别法 …… 39
5.3.1 基本思想…… 39
5.3.2 多元正态总体的 Bayes 判别法…… 40
5.4 费舍(Fisher)判别法 …… 42
5.4.1 两组判别分析…… 42
5.4.2 多组别费舍判别法…… 45
5.5 逐步判别法…… 46
5.5.1 引入和剔除变量所用的检验统计量…… 46
5.5.2 逐步判别的原则…… 47

第六章 主成分分析 …… 49

6.1 主成分分析的基本原理…… 49
6.2 主成分分析的推导…… 50
6.2.1 从协方差出发求解总体主成分…… 50
6.2.2 从相关阵出发求解总体主成分…… 52
6.2.3 样本的主成分…… 54

第七章 因子分析 …… 59

7.1 因子分析的基本理论…… 59
7.1.1 因子分析的数学模型…… 59
7.1.2 因子模型中的几个统计特征…… 60
7.2 因子载荷阵的估计方法…… 61
7.3 因子旋转…… 63
7.4 因子得分…… 65
7.5 因子分析的步骤与逻辑框图…… 66
7.5.1 因子分析的步骤…… 66
7.5.2 因子分析的逻辑框图…… 67

第八章　典型相关分析 …… 68

8.1　典型相关分析的数学描述 …… 69
8.2　总体典型相关 …… 69
8.3　样本典型相关 …… 72
8.4　典型相关系数的显著性检验 …… 73
8.5　典型相关系数的步骤及实例 …… 73

第九章　对应分析 …… 78

9.1　对应分析及基本思想 …… 78
9.1.1　对应分析的数据类型 …… 78
9.1.2　对应分析的基本思想 …… 80
9.2　列联表及列联表分析简介 …… 81
9.3　对应分析的基本理论 …… 82
9.3.1　距离与总惯量 …… 83
9.3.2　R 型与 Q 型因子分析的对等关系 …… 85
9.4　对应分析的步骤 …… 86

第十章　多维标度分析 …… 88

10.1　距离阵和经典解 …… 89
10.1.1　欧式距离阵 …… 89
10.1.2　欧式距离阵的判定定理 …… 89
10.1.3　多维标度的经典解 …… 91
10.2　实例 …… 91

第二篇　多元统计分析实验

实验一　均值向量和协方差阵的检验 …… 97

1.1　实验背景 …… 97
1.2　实验步骤和结果分析 …… 97

实验二　聚类分析 …… 109

2.1　实验背景 …… 109
2.2　实验步骤和结果分析 …… 109

实验三　判别分析 …… 121

3.1　实验背景 …… 121
3.2　实验步骤和结果分析 …… 121

实验四　主成分分析 …… 141

4.1　实验背景 …… 141
4.2　实验步骤和结果分析 …… 141

实验五　因子分析 …… 148

5.1　实验背景 …… 148
5.2　实验步骤和结果分析 …… 148

实验六　典型相关分析 …… 160

6.1　实验背景 …… 160
6.2　实验步骤和结果分析 …… 160

实验七　对应分析 …… 167

7.1　实验背景 …… 167
7.2　实验步骤和结果分析 …… 167

实验八　多维标度分析 …… 177

8.1　实验背景 …… 177
8.2　实验步骤和结果分析 …… 177

参考文献 …… 183

第一篇　多元统计分析原理与方法

第一章 绪 论

1.1 多元统计简介

在实际问题中，我们常常需要处理多个变量的观测数据。例如在研究公司的运营情况时，要考虑公司的获利能力、资金周转能力、竞争能力以及偿债能力等财务指标；又如衡量一个地区的经济发展水平，需要观察的指标有：总产值、利润、效益、劳动生产率等；在医学诊断中，有病还是无病，需做多项检测：血压、体温、心跳、白细胞等。显然，如果我们只研究一个指标或是将这些指标割裂开分别研究，是不能从整体上把握研究问题的实质的，解决这些问题就需要多元统计分析方法。多元统计分析(multivariate statistical analysis)就是研究多个随机变量之间相互依赖关系以及内在统计规律的一门学科，利用其中的不同方法可对研究对象进行分类和简化。

多元分析包括的主要内容有：多元正态总体的参数估计和假设检验以及常用的统计方法。这些方法有多元数据图表示法、聚类分析、判别分析、主成分分析、因子分析、对应分析、多维标度法、典型相关分析、路径分析、多重多元回归分析等。本书重点介绍前 8 种方法。

早在 19 世纪就出现了处理二维正态总体的一些方法，但系统地处理多维概率分布总体的统计分析问题，则开始于 20 世纪。人们常把 1928 年 Wishart 分布的导出作为多元分析成为一个独立学科的标志。20 世纪 30 年代，R. A. 费希尔、H. 霍特林、许宝禄以及 S. N. 罗伊等人作出了一系列奠基性的工作，使多元统计分析在理论上得到了迅速的进展。40 年代，多元分析在心理、教育、生物等方面获得了一些应用。由于应用时常需要大量的计算，加上第二次世界大战的影响，使其发展停滞了相当长的时间。50 年代中期，随着电子计算机的发展和普及，它在地质、气象、标准化、生物、图像处理、经济分析等许多领域得到了广泛的应用，也促进了理论的发展。20 世纪 60 年代通过应用和实践又完善和发展了理论，由于新的理论、新的方法不断涌现又促使它的应用范围更加扩大。70 年代初期在我国才受到各个领域的极大关注。近 40 多年来我国在多元统计方法的理论研究和应用上也取得了很多显著成绩，有些研究工作已达到国际水平，并已形成一支科技队伍，活跃在各条战线上。

1.2 主要内容安排

本书共分为十章。

第一章绪论，主要介绍多元分析研究对象及发展历史。第二章简要地介绍多元数据的

图表示法。第三章介绍多元分析的基本概念和基本理论。主要介绍多元正态总体的参数估计和假设检验。

第四章和第五章主要研究分类问题,介绍聚类分析法和判别分析法。实际应用时两种方法往往联合起来使用。因为判别分析要求对新样品进行判别分类之前,必先知道已有几类总体,然后建立判别式,对新样品进行判别归类。如果一批给出样品要划分几类事先不知道,这时可先做聚类分析然后再做判别分析。

第六章和第七章介绍主成分分析、因子分析。主要研究结构化简问题,将具有错综复杂关系的变量(或样品)综合成数量较少的因子尽可能简单地表示所研究的对象,又不至于损失很多有价值的信息。

第八章研究两组变量之间的相关关系,介绍典型相关分析,用于简化两组变量为少数综合变量以再现原来两组变量之间的相关关系。

第九章和第十章介绍了对应分析和多维标度分析,对应分析可以把变量点和样品点同时反映在同一个因子轴所确定的平面上(即取同一个坐标系),根据接近的程度,将变量点和样品点一起考虑进行分类。多维尺度分析通过低维空间(通常是二维空间)展示多个研究对象(样品)之间的联系,利用平面距离来反映研究对象之间的相似程度。这两种方法都是通过降维,在尽可能保留高维数据信息的前提下,把高维数据表达在平面图上,从而从视觉上简单的辨别出这些高维数据之间的关系。

本书除第一、二章之外,其余各章在统计方法介绍之后,都给出应用性课题的 SPSS 实现,共包含 8 个实验,供选作题参考,读者不妨就这些课题,收集有关数据,按每章所述方法去计算和分析,定有收获。本书的特点是将常用的多元分析方法的介绍与在计算机上实现这些方法的软件紧密地结合起来,不仅介绍每种多元分析方法的实际背景、统计思想、统计模型、数学原理和解题的思路,并结合实例介绍应用统计软件(Spss 系统)解决问题的步骤和计算结果的分析。

第二章　多元数据图表示法

图形是对资料进行探索性研究的重要工具，当变量较少时，可以采用直方图、条形图、饼图、散点图或是经验分布的密度图等方法。对于变量少于 3 个的情况这样做简单而有效。当变量个数大于 3 个时，就不能用通常的方法作图了。如果能把一些多元数据直接显示在平面图上，便可借助图形来描述多元数据的特性，从而使图形更加直观，简洁。自 20 世纪 70 年代以来，多元数据的图表示法一直是人们所关注的问题，期间涌现了很多方法，这些方法大体上分为两类：一类是使高维空间的点与二维、三维的空间上的某种图形对应，这种图形能反映高维数据的某些特点或数据间的某些关系；另一类是在尽可能多地保留原数据的信息的原则下进行降维，若使维数降低到三维以下，便可以在图形上直观的表达原数据的主要信息。后者主要的方法：主成分法，因子分析，对应分析，多维标度分析等。前者主要有散点图矩阵、雷达图，调和曲线图、脸谱图等，本章主要介绍这四种多变量的图表示法的基本思想及作图方法。

设指标（变量）数为 p，观测次数为 n（样品容量），第 α 次观测值记为 $X_\alpha=(x_{\alpha1},x_{\alpha2},\cdots,x_{\alpha p})'$，$\alpha=1,\cdots,n$。$n$ 次观测数据组成的矩阵为 $X=(x_{ij})_{n\times p}$

$$X=\begin{bmatrix}X_1'\\X_2'\\\vdots\\X_n'\end{bmatrix}=\begin{bmatrix}x_{11}&x_{12}&\cdots&x_{1p}\\x_{21}&x_{22}&\cdots&x_{2p}\\\vdots&\vdots&&\vdots\\x_{n1}&x_{n2}&\cdots&x_{np}\end{bmatrix}$$

例 1　考察 2011 年广东、江苏、陕西、甘肃四个省城镇居民家庭平均每人全年现金消费支出情况，选取五项指标，具体数据见表 2.1（摘自 2012 年中国统计年鉴）。

表 2-1　　（单位：元）

	肉禽及制品	住 房	医疗保健	交通和通信	文教和娱乐
广东	1926.38	541.63	948.18	3630.62	2647.94
江苏	1205.12	438	962.45	2262.19	2695.52
陕西	642.23	291.67	1100.51	1502.44	1857.6
甘肃	621.34	266.16	874.05	1289.8	1158.3

此例变量个数 $p=5$，观测次数 $n=4$。

2.1　散点图矩阵

散点图矩阵是借助两变量散点图的作图方法，它可以看作一个大的图形方阵，其每一个

非对角元素的位置上是对应行的变量与对应列的变量的散点图。它所研究的仍是两两变量之间的相关关系，而不能直接反映多个变量之间的关系，借助它来对资料分类也是比较困难的。

然而，因其直观、简单、容易理解，散点图矩阵还是越来越受到广大实际工作者的喜爱，很多统计软件也加入了作散点图矩阵的功能。

下面举例说明如何用 Spss 软件作散点图矩阵对资料进行研究，以 Spss 软件自带的 World95. sav 资料为例：

该资料共有 26 个变量、109 条观测，是 1995 年世界 109 个国家和地区的基本发展情况的资料。选择该亚洲地区的国家的几个变量作图：lifeexpf(女性预期寿命)、lifeexpm(男性预期寿命)、gdp_cap(GDP 是总资产的倍数)、birth_rt (婴儿出生率)，death_rt(婴儿死亡率)。

按 graphs→legacy dialogs→scatterplot/dot 顺序打开主对话框，matrix 为矩阵散点图(见图 2.1)。单击 define 按钮，展开 scatterplot matrix 矩阵散点图对话框，选择要分析的变量，点击 OK。

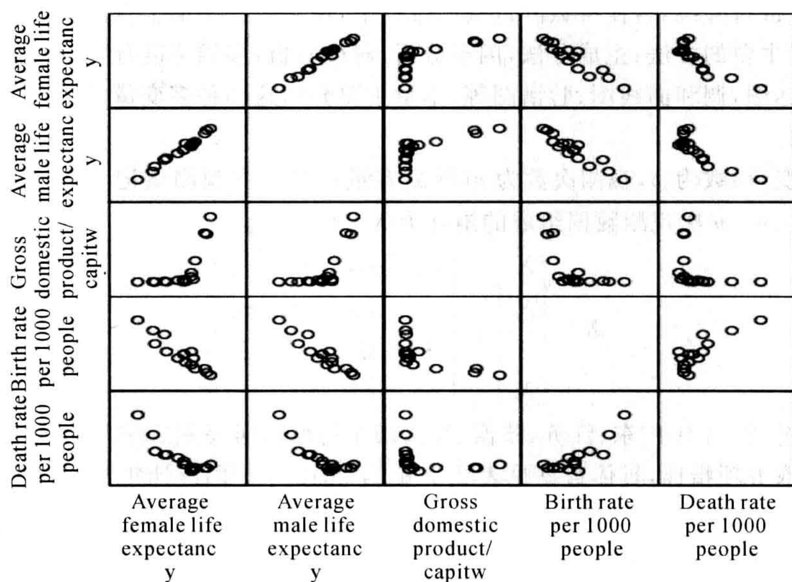


图 2.1

由散点图矩阵可以看到，男性的预期寿命、女性的预期寿命及婴儿死亡率，婴儿出生率四个变量之间有明显的线性相关性，而 GDP 是总资产的倍数与上面四个变量存在着某种曲线相关关系。

资料集 World95. sav 中变量 religion 的含义主要的宗教信仰，在 Sactterplot Matrix 对话框中将 regligion 变量选作标记变量(选入 Set markers by)，则在输出的散点图矩阵中，不同宗教信仰的国家以不同的颜色画出，这样可以做更详细的分析。

2.2 雷达图

雷达图是目前应用最为广泛的对多元资料进行作图的方法，利用雷达图可以很方便地研究各样本点之间的关系并进而对样品进行归类。设要分析的资料共有 p 个变量，雷达图的标准画法如下：

(1)作一圆，并把圆周分为 p 等分。

(2)连接圆心和各分点，把这 p 条半径依次定义为各变量的坐标轴，并标以适当的刻度。

(3)对给定的一次观测值，把它的 p 个分量值分别点在相应的坐标轴上，然后连接成一个 p 边形，这个 p 边形就是 p 元观测值的图示，n 次观测值可画出 n 个 p 边形。

Excel 软件提供了画雷达图的功能，它适合于观测数(样品)较少的情形，这时可以方便地把各观测画到一张图里面，便于对各指标进行对比，将例 1 数据用雷达图表示如图 2.2。

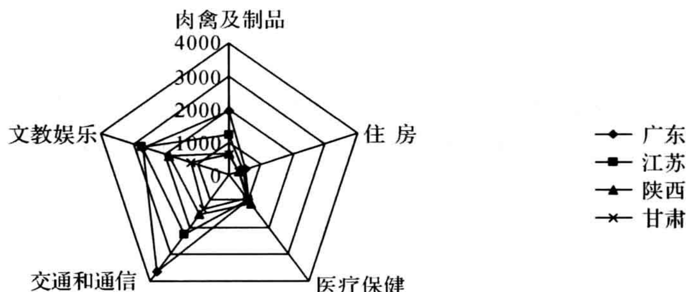


图 2.2

这种图形既像雷达荧光屏上看到的图像，也像个蜘蛛网，因此有人称为雷达图，也有人称为蛛网图。利用雷达图有助于观测多元数据的某些特点，便于进行分析，例如从上图不难看出广东、江苏各种指标都较高，对应着一个面积较大的五边形。而陕西、甘肃各种指标都较低，其图形面积也较小，利用图形和面积大小可对样品进行初始分类，将广东、江苏分为一类，陕西、甘肃分为一类。

当观测数比较多时，画到一张雷达图里面就不太容易看出各观测之间的接近程度，用 Excel 当然也可以对每一个观测画一张雷达图。值得注意的是，这里坐标轴只有正半轴，因而只能表示非负数据，若有负数据，只能通过合理变换使之非负才行。

2.3 调和曲线图

调和曲线图是 D. F. Andrews1972 年提出的三角多项式作图法，所以又称为三角多项式图，其思想是把高维空间中的一个样品点对应于二维平面上的一条曲线。

设 p 维数据 $X=(x_1,x_2,\cdots,x_p)'$ 对应的曲线是

$$f_x(t)=\frac{x_1}{\sqrt{2}}+x_2\sin t+x_3\cos t+x_4\sin 2t+x_5\cos 2t+\cdots$$

$$-\pi\leqslant t\leqslant 5$$

上式当 t 在区间$(-\pi,\pi)$上变化时，其轨迹是一条曲线。

n 次观测对应 n 条曲线，画在同一平面上就是一张调和曲线图。在多项式的图表示中，当各变量的数值太悬殊时，最好先标准化后再作图。

从数学上看，调和曲线图是一种较好的图示法，因为它具有许多好的性质，

(1)保线性关系

设 X,Y,Z 均为 p 维向量，a,b 为常数，若 $Z=aX+bY$ 则

$$f_z(t)=af_X(t)+bf_Y(t)\ ,\ -\pi\leqslant t\leqslant\pi$$

(2)保欧氏距离

由于 $f_X(t)f_Y(t)$都是$[-\pi,\pi]$上的平方可积函数，定义它们之间的欧氏距离为：

$$d^2_{f_Xf_Y}=\int_{-\pi}^{\pi}|f_X(t)-f_Y(t)|^2\mathrm{d}t$$

则它与 X,Y 的欧氏距离 $d^2_{XY}=(X-Y)'(X-Y)$ 有关系：

$$d^2_{XY}=\frac{1}{\pi}d^2_{f_Xf_Y}$$

这就是说原来两个样品之间的欧氏距离与变换后两条曲线的距离只差一个倍数，故调和曲线图对聚类分析帮助很大，同类的曲线非常靠近拧在一起，不同类的曲线相互分开，非常直观。

作调和曲线时一般要借助计算机作图，我们利用 Matlab 画出例 1 的数据所代表的调和曲线图：

```
X=[1926.38    541.63    948.18    3630.62    2647.94;
   1205.12    438    962.45    2262.19    2695.52;
   642.23    291.67    1100.51    1502.44    1857.6;
   621.34    266.16    874.05    1289.8    1158.3]
   species={'广东','江西','陕西','甘肃'}
   andrewsplot(X,'group',species)
```

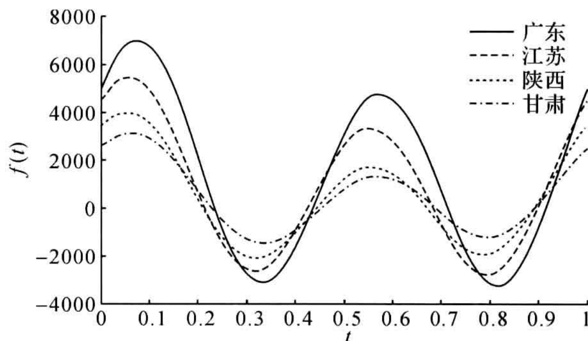


图 2.3

从图 2.3 可以看出，广东、江苏可以归为一类，陕西、甘肃归为一类。

2.4 脸谱图

脸谱图用脸谱来表达多变量的样品，由美国统计学家 H. Chernoff 于 1970 年首先提出，该方法是将观测的各个变量（指标）分别用脸的某一部位的形状或大小来表示，一个样品（观测）可以画成一张脸谱。他首先将该方法用于聚类分析，引起了各国统计学家的极大兴趣，并对他的画法作出了改进。

按照切尔诺夫于 1973 年提出的画法，脸谱图采用 15 个指标，各指标代表的面部特征为：1 表示脸的范围，2 表示脸的形状，3 表示鼻子的长度，4 表示嘴的位置，5 表示笑容曲线，6 表示嘴的宽度，7—11 分别表示眼睛的位置，分开程度，角度，形状和宽度，12 表示瞳孔的位置，13—15 分别表示眼眉的位置，角度及宽度。这样，按照各变量的取值，根据一定的数学函数关系，就可以确定脸的轮廓、形状及五官的部位、形状，每一个样本点都用一张脸谱来表示。而脸谱容易给人们留下较为深刻的印象，通过对脸谱的分析，就可以直观地对原始资料进行归类或比较研究。

在实际问题的分析中，如果数据的指标数小于 15，则脸部有些特征将被自动固定。统计学曾给出了几种不同的脸谱图的画法，而对于同一种脸谱图的画法，将变量次序重新排列，得到的脸谱的形状也会有很大不同。此处我们不对脸谱的各个部位与原始变量的数学关系作过多探讨，而只说明其作图的思想及软件实现方法。

Matlab 软件收录了脸谱图的作图方法，这里先介绍 Matlab 中实现脸谱图的一个基本函数：glyphplot(X,‘glyph’,‘face’,‘features’,f)

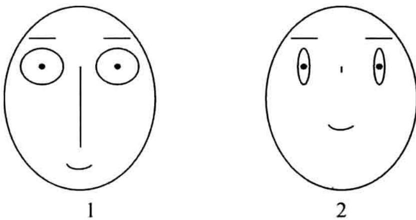
在这个函数中，X 代表样本矩阵，f 定义 X 中各行（样本）由脸部的那一部分特征表示，下面的表格是 Matlab 的脸部特征所对应的代码：

Column	Facial Feature
1	Size of face
2	Forehead/jaw relative arc length
3	Shape of forehead
4	Shape of jaw
5	Width between eyes
6	Vertical position of eyes
7	Height of eyes
8	Width of eyes (this also affects eyebrow width)
9	Angle of eyes (this also affects eyebrow angle)
10	Vertical position of eyebrows
11	Width of eyebrows (relative to eyes)
12	Angle of eyebrows (relative to eyes)
13	Direction of pupils
14	Length of nose
15	Vertical position of mouth
16	Shape of mouth
17	Mouth arc length

下面我们举例说明如何用 Matlab 软件画脸谱图，比如，

```
X=[2.89,5.16;1,4.89]
glyphplot(X,'glyph','face','features',[14,7])
```

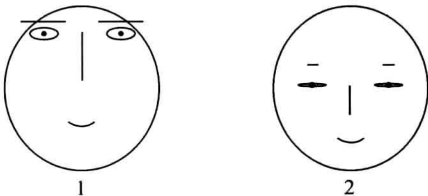
结果：



上边的数据 $X=\begin{bmatrix} 2.89 & 5.16 \\ 1 & 4.89 \end{bmatrix}$，有两个指标(列数)，第一个指标对应的脸部特征是"14"，表示鼻子的长度，第二指标对应的脸部特征是"7"，表示眼睛的高度。经过定义以后，除了眼睛与鼻子之外，两个样本输出图形的其他脸部特征是相同的。在本例中，样本1的鼻子长，表明样本1的第一个指标比样本2的第一个指标大得多，同样的，样本1的眼睛比样本2的眼睛大，说明样本1的第二指标比样本2的第二指标大。对于相同的数据，我们可以改变样品对应的脸部特征的定义：

```
glyphplot(X,'glyph','face','features',[6,11])
```

可得图形：



最后，让我们再来看一个实例。

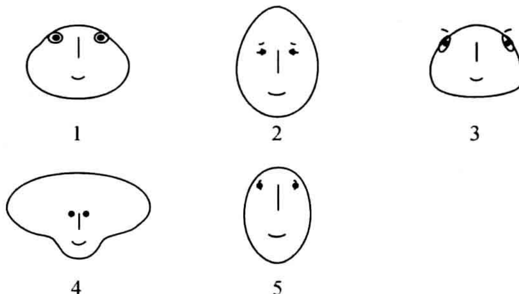
例3 下表是五大钢铁公司反映经营状况的十大指标，为了比较国内钢铁公司与韩国蒲项钢铁公司的差距，下面做出韩国蒲项钢铁公司、宝钢、鞍钢、武钢、首钢五家钢铁公司的脸谱图。

项目	宝钢	鞍钢	武钢	首钢	浦钢
负债保障率	2.89	2.95	2.34	1.85	3.12
长期负债倍数	5.16	9.15	6.07	2.63	6.96
流动比率	1.31	1.83	1.16	2.22	2.1
资产利润率	21.71	17.34	24.77	11.89	25.34
收入利润率	23.17	11.33	19.55	7.6	22.28
成本费用利率	30.23	12.76	24.81	8.06	28.52
净利润现金比率	1.79	0.9	1.7	1.09	1.3
三年资产平均增长率	1.48	7.28	63.3	11.76	13.18
三年销售平均增长率	20.07	29.19	52.88	18.77	24.16
三年平均资本增长率	11.04	10.5	48.95	7.63	17.51

Matlab 中输入数据：

```
A=[2.89,5.16,1.31,21.71,23.17,30.23,1.79,1.48,20.07,11.04;
   2.95,9.15,1.83,17.34,11.33,12.76,0.9,7.28,29.19,10.5;
   2.34,6.07,1.16,24.77,19.55,24.81,1.7,63.3,52.88,48.95;
   1.85,2.63,2.22,11.89,7.6,8.05,1.09,11.76,18.77,7.63;
   3.12,6.96,2.1,25.34,22.28,28.52,1.3,13.18,24.16,17.51]
glyphplot(A,'glyph','face','features',[2,3,4,5,6,7,8,9,10,11]);
```

可得：



可按照特征表进行各个指标的解释，也能从总体上提供分类的依据，上图形看，5 个样品可以分为三类，样本 2,5 可以分成一类，1,3 分成一类，4 分成一类。

本章思考与练习

1. 试述多变量图示法的方法思想和实际意义。
2. 散点图，雷达图，调和曲线图，脸谱图适合的场合及特点是什么？
3. 以下是两家上市公司某年的部分收益性及成长性财务指标：

公司简称	深能源 A	深南电 A
净资产收益率/%	16.85	22
总资产报酬率/%	12.35	15.3
资产负债率/%	42.32	46.51
总资产周转率/%	0.37	0.76
流动资产周转率/%	1.78	1.77
已获利息倍数/%	7.18	15.67
销售增长率/%	45.73	48.11
资本积累率/%	54.54	19.41

试用本章所学的图形描述上述数据，并做简单的分析。

第三章　均值向量和协方差阵的检验

多元统计分析涉及的都是随机向量或随机向量放在一起组成的随机矩阵，由于随机向量的多元正态分布在多元统计分析的理论和实际运用中都有着重要的地位，本章着重介绍多元正态分布的定义、参数估计及相关的检验。在介绍正态分布之前，本章先介绍有关随机向量的基本概念。

本章的不少内容是一元的直接推广，但由于多指标问题的复杂性，在这里将只列出检验用的统计量，介绍如何使用这些统计量，而对有关的检验问题的理论推导全部略去。

3.1　随机向量

这里研究的对象是多个变量的总体，即同时收集了 p 个指标(变量)，又进行了 n 次观测得到观测样品(样本)，我们把这个 p 指标表示为 $X_1,X_2,\cdots,X_p$，常用向量 $X=(X_1,X_2,\cdots,X_p)'$表示对同一个体观测的 p 个变量，在这一节中，我们将介绍随机向量的基本概念。

定义 3.1　将 p 个随机变量 $X_1,X_2,\cdots,X_p$ 的整体称为 p 维随机向量，记为 $X=(X_1,X_2,\cdots,X_p)'$。

定义 3.2　设 $X=(X_1,X_2,\cdots,X_p)'$是 p 维随机向量，它的多元分布函数定义为

$F(x)\triangleq F(x_1,x_2,\cdots,x_p)=P(X_1\leqslant x_1,X_2\leqslant x_2,\cdots,X_p\leqslant x_p)$。

记 $X\sim F(x)$，其中 $x=(x_1,x_2,\cdots,x_p)'\in R^p$，$R^p$ 表示 p 维欧氏空间。

定义 3.3　设 $X=(X_1,X_2,\cdots,X_p)'$是 p 维随机向量，若存在有限个或可列个 p 维数向量 $x_{(1)},x_{(2)},\cdots$，记 $P(X=x_{(k)})=p_k$，且满足 $p_1+p_2+\cdots=1$，则称 X 为离散型随机向量，称 $P(X=x_{(k)})=p_k$，为 X 的概率分布。

定义 3.4　设 $X\sim F(x)\triangleq F(x_1,x_2,\cdots x_p)$，若存在非负函数 $f(x_1,x_2,\cdots,x_p)$，使得

$$F(x)\triangleq F(x_1,x_2,\cdots x_p)=\int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_p}f(t_1,t_2,\cdots,t_p)\mathrm{d}t_1\cdots\mathrm{d}t_p$$

则称 X 为连续型随机变量，$f(x_1,x_2,\cdots x_p)$为分布密度函数。

定义 3.5　设 $X=(X_1,X_2,\cdots,X_p)'$，若 $E(X_i)(i=1,\cdots,p)$存在且有限，则称 $E(X)=(E(X_1),E(X_2),\cdots,E(X_p))'$为 X 的均值(向量)或数学期望，有时也把 $E(X)$和 $E(X_i)$分别记为 μ 和 μ_i，即 $\mu=(\mu_1,\mu_2,\cdots,\mu_p)'$。

容易推得均值(向量)具有以下性质：

(1)$E(AX)=AE(X)$

(2)$E(AXB)=AE(X)B$

(3)$E(AX+BY)=AE(X)+BE(Y)$

其中，X、Y 为随机向量，A、B 为大小适合运算的常数矩阵。

定义 3.6 设 $X=(X_1,X_2,\cdots,X_p)'$，$Y=(Y_1,Y_2,\cdots,Y_p)'$，称

$$D(X)\triangleq E(X-E(X))(X-E(X))'=\begin{bmatrix}\mathrm{Cov}(X_1,X_1) & \mathrm{Cov}(X_1,X_2) & \cdots & \mathrm{Cov}(X_1,X_p)\\ \mathrm{Cov}(X_2,X_1) & \mathrm{Cov}(X_2,X_2) & \cdots & \mathrm{Cov}(X_2,X_p)\\ \vdots & \vdots & & \vdots\\ \mathrm{Cov}(X_p,X_1) & \mathrm{Cov}(X_p,X_2) & \cdots & \mathrm{Cov}(X_p,X_p)\end{bmatrix}$$

为 X 的方差或协差阵，有时把 $D(X)$ 简记为 $\sum$，$\mathrm{Cov}(X_i,X_j)$ 简记为 σ_{ij}，从而有

$$\sum=(\sigma_{ij})_{p\times p}。$$

称随机向量 X 和 Y 的协差阵为：

$$\mathrm{Cov}(X,Y)\triangleq E(X-E(X))(Y-E(Y))'$$

$$=\begin{bmatrix}\mathrm{Cov}(X_1,Y_1) & \mathrm{Cov}(X_1,Y_2) & \cdots & \mathrm{Cov}(X_1,Y_p)\\ \mathrm{Cov}(X_2,Y_1) & \mathrm{Cov}(X_2,Y_2) & \cdots & \mathrm{Cov}(X_2,Y_p)\\ \vdots & \vdots & & \vdots\\ \mathrm{Cov}(X_p,Y_1) & \mathrm{Cov}(X_p,Y_2) & \cdots & \mathrm{Cov}(X_p,Y_p)\end{bmatrix}$$

当 $X=Y$ 时，即为 $D(X)$。若 $\mathrm{Cov}(X,Y)=0$，则称 X 和 Y 不相关，由 X 和 Y 相互独立易推得 $\mathrm{Cov}(X,Y)=0$，即 X 和 Y 不相关；但反过来，当 X 和 Y 不相关时，一般不能推知它们独立。

若 $X=(X_1,X_2,\cdots,X_p)'$ 的协差阵存在，且每个分量的方差大于零，则称随机向量 X 的相关阵为：$R=(r_{ij})_{p\times p}$，其中

$$r_{ij}=\frac{\mathrm{Cov}(X_i,X_j)}{\sqrt{D(X_i)}\ \sqrt{D(X_j)}}=\frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\ \sqrt{\sigma_{jj}}}$$

为相关系数。

当 A、B 为常数矩阵时，由定义可以推出协差阵有如下性质：

(1) 对于常数向量 a，有 $D(X+a)=D(X)$

(2) $D(AX)=AD(X)A'=A\sum A'$

(3) $\mathrm{Cov}(AX,AY)=A\mathrm{Cov}(X,Y)B'$

最后，我们应该注意到，对于任何的随机向量 $X=(X_1,X_2,\cdots,X_p)'$ 来说，其协差阵 $\sum$ 都是对称阵，同时总是非负定(半正定)的，大多数情况是正定的。

3.2 多元正态分布

多元正态分布在多元统计分析中所占的重要地位，如同一元统计分析中一元正态分布所占的重要地位一样，多元统计分析的许多重要理论和方法都是直接或间接建立在正态分布的基础上，多元正态分布是多元统计分析的基础。

3.2.1 多元正态分布的定义及基本性质

多元正态分布有多种定义方法，下面给出最常用的一种，并列出其相关的性质。

定义 3.7 若 p 维随机向量 $X=(X_1,\cdots,X_p)'$ 的密度函数为：

$$f(x_1,\cdots,x_p)=\frac{1}{\sqrt{(2\pi)^p}\mid\sum\mid^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu)'\sum^{-1}(x-\mu)\right\}$$

其中 $x=(x_1,\cdots,x_p)'$，μ 是 p 维向量，$\sum$ 是 p 阶正定阵，则称 X 为 p 维正态随机向量，

简记：$X\sim N_p(\mu,\sum)$。可以证明 μ 为 X 的均值向量 $\sum$，为 X 的协差阵。

多元正态变量的基本性质：

(1) 若 $X=(X_1,X_2,\cdots,X_p)'\sim N_p(\mu,\sum)$，$\sum$ 是对角阵，则 $X_1,\cdots X_p$ 相互独立

(2) 若 $X=(X_1,X_2,\cdots,X_p)'\sim N_p(\mu,\sum)$，$A$ 为 $s\times p$ 阶常数阵，d 为 s 维常数向量，则：$AX+d\sim N_s(A\mu+d,A\sum A')$

(3) 若 $X=(X_1,X_2,\cdots,X_p)'\sim N_P(\mu,\sum)$，将 $X,\mu,\sum$ 作如下剖分：

$$X=\begin{bmatrix}X^{(1)}\\X^{(2)}\end{bmatrix}\begin{matrix}q\\p-q\end{matrix},\mu=\begin{bmatrix}\mu^{(1)}\\\mu^{(2)}\end{bmatrix}\begin{matrix}q\\p-q\end{matrix},\sum=\begin{bmatrix}\sum_{11}&\sum_{12}\\\sum_{21}&\sum_{22}\end{bmatrix}\begin{matrix}q\\p-q\end{matrix}$$

则 $X^{(1)}\sim N_q(\mu^{(1)},\sum_{11})$，$X^{(2)}\sim N_{p-q}(\mu^{(2)},\sum_{22})$。

3.2.2 多元正态分布的参数估计

(一) 多元样本的概念及表示法

多元分析研究的总体是多元总体，设多元总体 $X\sim N_p(\mu,\Sigma)$，随机抽取 n 个个体：$X_{(1)}$，$X_{(2)}$，$\cdots X_{(n)}$，若其满足下列两个条件：

(1) $X_{(1)},X_{(2)},\cdots X_{(n)}$ 相互独立

(2) $X_{(1)},X_{(2)},\cdots X_{(n)}$ 与总体同分布

则称 $X_{(1)},X_{(2)},\cdots X_{(n)}$ 为该总体的一个多元随机样本，简称简单样本。

每个 $X_{(a)}=(X_{a1},X_{a2},\cdots,X_{ap})'(a=1,2,\cdots,n)$ 为一个样品(样本)，将全部观测结果用一个 $n\times p$ 阶矩阵

$$X=\begin{bmatrix}X'_{(1)}\\X'_{(2)}\\\vdots\\X'_{(n)}\end{bmatrix}=\begin{bmatrix}X_{11}&X_{12}&\cdots&X_{1p}\\X_{21}&X_{22}&\cdots&X_{2p}\\\vdots&\vdots&&\vdots\\X_{n1}&X_{n2}&\cdots&X_{np}\end{bmatrix}$$

称 X 为样本资料矩阵。

值得注意的是：

(1) 多元样本中的每个样品，对 p 个指标的观测值往往是有相关关系的，但不同样品之间的观测值一定是相互独立的。

(2) 多元分析所处理的多元样本观测数据一般都属于横截面数据，即在同一时间不同空间上的数据。

(二) μ 和 $\sum$ 的极大似然估计

多元总体 $X\sim N_p(\mu,\Sigma)$，$X_{(1)},X_{(2)},\cdots X_{(n)}$ 为该总体的一个多元随机样本

(1) 样本均值向量：$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_{(i)} = \frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n} X_{i1} \\ \sum_{i=1}^{n} X_{i2} \\ \vdots \\ \sum_{i=1}^{n} X_{ip} \end{bmatrix} = \begin{bmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \vdots \\ \overline{X}_p \end{bmatrix}$

(2) 样本离差阵：$S = \sum_{i=1}^{n}(X_{(i)} - \overline{X})(X_{(i)} - \overline{X})' = (s_{ij})_{p\times p}$

$$S_{p\times p} = \begin{bmatrix} \sum_{i=1}^{n}(X_{i1} - \overline{X}_1)^2 & \cdots & \sum_{i=1}^{n}(X_{i1} - \overline{X}_1)(X_{ip} - \overline{X}_{ip}) \\ & \sum_{i=1}^{n}(X_{i2} - \overline{X}_2)^2 \quad \cdots & \sum_{i=1}^{n}(X_{i2} - \overline{X}_2)(X_{ip} - \overline{X}_{ip}) \\ & & \vdots \\ & & \sum_{i=1}^{n}(X_{ip} - \overline{X}_p)^2 \end{bmatrix}$$

(3) 样本协差阵定义为：$V_{p\times p} = \frac{1}{n}S = \frac{1}{n}\sum_{a=1}^{n}(X_{(a)} - \overline{X})(X_{(a)} - \overline{X})' = (v_{ij})_{p\times p}$

(4) 样本相关阵定义为：$R = (r_{ij})_{p\times p}$，其中 $r_{ij} = \frac{v_{ij}}{\sqrt{v_{ii}}\sqrt{v_{jj}}} = \frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}}$

可以证明 μ 和 $\sum$ 的极大似然估计为：

$$\hat{\mu} = \overline{X}, \hat{\sum} = \frac{1}{n}S$$

3.3 均值向量的检验

在单一变量的统计分析中，已经给出了正态总体 $N(\mu,\sigma^2)$ 的均值 μ 和方差 σ^2 的各种检验。对于多变量的正态总体 $N_p(\mu,\sum)$，各种实际问题同样要求对 μ 和 $\sum$ 进行统计推断。例如，要考察全国各省、自治区和直辖市的社会经济发展状况，与全国平均水平相比较有无显著性差异等，就涉及多元正态总体均值向量的检验问题等。本章类似单一变量统计分析中的各种均值和方差的检验，相应地给出多元统计分析中的各种均值向量和协差阵的检验。

3.3.1 单个正态总体 $N_p(\mu,\sum)$ 均值向量的检验

设 p 元正态总体 $N_p(\mu,\sum)$，从总体中抽取容量为 n 的样本：

$$X_{(1)}, X_{(2)}, \cdots, X_{(n)}, \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_{(i)}, S = \sum_{i=1}^{n}(X_{(i)} - \overline{X})(X_{(i)} - \overline{X})'$$

(1) 已知总体协方差阵 $a_{1i}^2 + a_{2i}^2 + \cdots + a_{pi}^2 = 1(i = 1,2,\cdots,m)$

$$H_0: \mu = \mu_0(\mu_0\text{ 为已知均值向量}) \qquad H_1: \mu \neq \mu_0$$

在 H_0 成立的条件下，检验统计量 $T_0^2 = n(\overline{X} - \mu_0)' \sum^{-1} (\overline{X} - \mu_0) \sim \chi^2(p)$，

拒绝域：$W = \{T_0^2 > \chi_\alpha^2(p)\}$。

(2) 未知总体协方差阵 $\sum$

$H_0: \mu = \mu_0(\mu_0$ 为已知均值向量) $\qquad H_1: \mu \neq \mu_0$

在 H_0 成立的条件下，检验统计量 $\dfrac{n-p}{(n-1)p} T^2 \sim F(p, n-p)$

其中 $T^2 = (n-1)[\sqrt{n}(\overline{X} - \mu_0)' S^{-1} \sqrt{n}(\overline{X} - \mu_0)]$

拒绝域：$W = \{\dfrac{(n-1)p}{(n-p)} T^2 > F_\alpha(p, n-p)\}$。

3.3.2 两个正态总体 $N_p(\mu_1, \sum_1)$ 和 $N_p(\mu_2, \sum_2)$ 均值向量的检验

设

$$X_{(\alpha)} = (X_{\alpha 1}, X_{\alpha 2}, \cdots, X_{\alpha p})' \sim N_p(\mu_1, \Sigma_1) \quad \alpha = 1, \cdots, n$$
$$Y_{(\alpha)} = (Y_{\alpha 1}, Y_{\alpha 2}, \cdots, Y_{\alpha p})' \sim N_p(\mu_2, \Sigma_2) \quad \alpha = 1, \cdots, m$$

(1) 有共同已知协差阵 $\sum_1 = \sum_2 = \sum$

$$H_0: \mu_1 = \mu_2, \qquad H_1: \mu_1 \neq \mu_2$$

在 H_0 成立的条件下，检验统计量 $T_0^2 = \dfrac{nm}{n+m} (\overline{X} - \overline{Y})' \sum^{-1} (\overline{X} - \overline{Y}) \sim \chi^2(p)$

拒绝域：$W = \{T_0^2 > \chi_\alpha^2(p)\}$。

(2) 有共同的未知协差阵 $\sum > 0$

$$H_0: \mu_1 = \mu_2, \qquad H_1: \mu_1 \neq \mu_2 \quad ,$$

在 H_0 成立的条件下，检验统计量

$$F = \frac{(n+m-2)-p+1}{(n+m-2)p} T^2 \sim F(p, n+m-p+1)$$

其中：

$$T^2 = (n+m-2)\left[\sqrt{\frac{nm}{n+m}}(\overline{X} - \overline{Y})\right]' S^{-1} \left[\sqrt{\frac{nm}{n+m}}(\overline{X} - \overline{Y})\right]$$
$$S = S_1 + S_2$$
$$S_1 = \sum_{\alpha=1}^{n} (X_{(\alpha)} - \overline{X})(X_{(\alpha)} - \overline{X})'$$
$$S_2 = \sum_{\alpha=1}^{m} (Y_{(\alpha)} - \overline{Y})(Y_{(\alpha)} - \overline{Y})'$$

拒绝域：$W = \{F > F_\alpha(p, n+m-p-1)\}$。

(3) 当协方差阵不相等，则 $\sum_1 \neq \sum_2$ 且 $\sum_1 > 0, \sum_2 > 0$ 时

$$H_0: \mu_1 = \mu_2, \; H_1: \mu_1 \neq \mu_2$$

$n = m$

令

$$Z_{(i)} = X_{(i)} - Y_{(i)},\ i = 1,2,\cdots,n$$

$$\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_{(i)} = \overline{X} - \overline{Y}$$

$$S = \sum_{i=1}^{n}(Z_{(i)} - \overline{Z})(Z_{(i)} - \overline{Z})'$$

在 H_0 成立的条件下，检验统计量

$$F = \frac{(n-p)n}{p}\overline{Z}'S^{-1}\overline{Z} \sim F(p,n-p)$$

拒绝域：$W = \{F > F_\alpha(p,n-p)\}$。

$n \neq m$

在此，我们不妨假设 $n < m$，令

$$Z_{(i)} = X_{(i)} - \sqrt{\frac{n}{m}}Y_{(i)} + \frac{1}{\sqrt{nm}}\sum_{j=1}^{n} Y_{(j)} - \frac{1}{m}\sum_{j=1}^{m} Y_{(j)},\ i = 1,2\cdots,n$$

$$\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_{(i)} = \overline{X} - \overline{Y}$$

$$S = \sum_{i=1}^{n}(Z_{(i)} - \overline{Z})(Z_{(i)} - \overline{Z})'$$

假设 H_0 成立时，检验统计量为

$$F = \frac{(n-p)n}{p}\overline{Z}'S^{-1}\overline{Z} \sim F(p,n-p)$$

拒绝域：$W = \{F > F_\alpha(p,n-p)\}$

3.3.3 多个正态总体均值向量的检验（多元方差分析）

设有 k 个 p 元正态总体 $N_p(\mu_1,\Sigma),\cdots,N_p(\mu_p,\Sigma)$，从每个总体抽取独立样品个数为 n_1，$n_2,\cdots,n_k,n_1+\cdots+n_k \triangleq n$。欲检验假设：

$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$ $\quad$ $H_0: \mu_1,\mu_2,\cdots,\mu_k$ 至少有两个不相等。

每个样品观测 p 个指标得观测数据如下：

第一个总体：$X_i^{(1)} = (X_{i1}^{(1)},X_{i2}^{(1)},\cdots,X_{ip}^{(1)}),i = 1,2,\cdots,n_1$，

第二个总体：$X_i^{(2)} = (X_{i1}^{(2)},X_{i2}^{(2)},\cdots,X_{ip}^{(2)}),i = 1,2,\cdots,n_2$，

…… …… ……

第 k 个总体：$X_i^{(k)} = (X_{i1}^{(k)},X_{i2}^{(k)},\cdots,X_{ip}^{(k)}),i = 1,2,\cdots,n_k$

全部样品的总均值向量：

$$\overline{X} = \frac{1}{n}\sum_{\alpha=1}^{k}\sum_{i=1}^{n_\alpha} X_{(i)}^{(\alpha)} \triangleq (\overline{X}_1,\overline{X}_2,\cdots,\overline{X}_p)$$

各总体样品的均值向量：

$$\overline{X}^{(\alpha)} = \frac{1}{n_\alpha}\sum_{i=1}^{n_\alpha} X_{(i)}^{(\alpha)} \triangleq (\overline{X}_1^{(\alpha)},\overline{X}_2^{(\alpha)},\cdots,\overline{X}_p^{(\alpha)}),\alpha = 1,\cdots,k$$

组间离差阵： $A = \sum_{\alpha=1}^{k} n_\alpha(\overline{X}^{(\alpha)} - \overline{X})(\overline{X}^{(\alpha)} - \overline{X})'$

组内离差阵：$E = \sum_{\alpha=1}^{k}\sum_{i=1}^{n_\alpha}(X_{(i)}^{(\alpha)} - \overline{X}^{(\alpha)})(X_{(i)}^{(\alpha)} - \overline{X}^{(\alpha)})'$

总离差阵： $T = \sum_{a=1}^{k}\sum_{i=1}^{n_a}(X_{(i)}^{(a)} - \overline{X})(X_{(i)}^{(a)} - \overline{X})'$

$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$　　$H_0: \mu_1, \mu_2, \cdots, \mu_k$ 至少有两个不相等

在 $f = \frac{1}{2}p(p+1)(k-1)$ 成立时，检验统计量 $\Lambda = \frac{|E|}{|T|} = \frac{|E|}{|A+E|} \sim \Lambda(p, n-k, k-1)$（Wilks 分布）。

在这里我们特别要注意，Wilks 分布表也可用 λ_k 分布或 λ'_k 分布来近似，巴特莱特(Bartlett) 提出了用 $-2\ln\lambda'_k$ 分布来近似。设 $\Lambda \sim \Lambda(p, n-k, k-1)$，令

$$V = -(n-1-(p+k)/2)\ln\Lambda = \ln\Lambda^{-t}$$

则 $i = 1, \cdots, k$ 近似服从 $\chi^2(p(k-1))$ 分布。其中，$t = n-1-(p+k)/2$。

Rao 后来又研究用 n_i 分布来近似. 设 $\Lambda \sim \Lambda(p, n-k, k-1)$，令

$$R = \frac{1-\Lambda^{1/L}}{\Lambda^{1/L}} \cdot \frac{tL-2\lambda}{p(k-1)}.$$

则 $n_i - 1$ 近似服从 $F(p(k-1), tL-2\lambda)$，这里 λ'_k 不一定为整数，可用与它最近的整数来作为 $-2\ln\lambda'_k$ 的自由度，且 $\min(p, k-1) > 2$. 其中，$t = n-1-(p+k)/2$，$L = \left(\frac{p^2(k-1)^2-4}{p^2+(k-1)^2-5}\right)^{1/2}$，$\lambda = \frac{p(k-1)-2}{4}$。

3.4　协差阵的检验

3.4.1　一个正态总体协方差阵的检验

设 $X_{(\alpha)} = (X_{\alpha 1}, X_{\alpha 2}, \cdots, X_{\alpha p})'$ $\alpha = 1, \cdots, n$ 为来自 p 维正态总体 $N_p(\mu, \sum)$ 的样本，$\sum$ 未知，且 $\sum > 0$。

(1) $H_0: \sum = I_p$，　$H_1: \sum \neq I_p$

所构造的检验统计量为　$\lambda = \exp\left\{-\frac{1}{2}trS\right\}\left|S\right|^{\frac{n}{2}}\left(\frac{e}{n}\right)^{\frac{np}{2}}$

其中 S 为样本离差阵。在 H_0 成立时，$-2\ln\lambda$ 极限分布是 $\chi^2(p(p+1)/2)$ 分布，因此当 $n \gg p$，由样本值计算出 λ 值，若 $-2\ln\lambda > \chi^2_\alpha(p(p+1)/2)$，即 $\lambda < e^{-\frac{\chi^2_\alpha(p(p+1)/2)}{2}}$，则拒绝 H_0，否则接受 H_0。

(2) $H_0: \sum = \sum_0 \neq I_p$，　$H_1: \sum \neq \sum_0 \neq I_p$

因为 $\sum_0 > 0$，所以存在 $D(|D| \neq 0)$，使得 $D\sum_0 D' = I_p$

令 $Y_{(a)} = DX_{(a)}$，$a = 1, 2, \cdots, n$，

则 $Y_{(a)} \sim N_p(D\mu, D\sum D') = N_p(\mu^*, \sum^*)$，

因此，检验 $\sum = \sum_0$ 等价于检验 $\sum^* = I_p$。

3.4.2　多个正态总体协方差阵的检验

设有 k 个正态总体分别为 $N_p(\mu_1, \sum_1)$，$f = \frac{1}{2}p(p+1)(k-1)N_p(\mu_2, \sum_2)$，$\sum_i > 0$

且未知，$i=1,\cdots,k$。从 k 个总体分别取 n_i 个样本

$$X_{(a)}^{(i)}=(X_{a1}^{(i)},\cdots,X_{ap}^{(i)})',a=1,\cdots,n_i,i=1,\cdots,k$$

这里 $\sum_{i=1}^{k}n_i=n$ 为总样本容量。

$H_0:\sum_1=\sum_2=\cdots=\sum_k$，　$H_1:\sum_1,\sum_2,\cdots,\sum_k$ 至少有两个不相等

构造统计量如下：

$$\begin{aligned}-2\ln\lambda'_k&=(n-k)\ln\left|\frac{1}{n-k}S\right|-\sum_{i=1}^{k}(n_i-1)\ln\left|\frac{1}{n_i-1}S_i\right|\\&=(n-k)\ln|V|-\sum_{i=1}^{k}(n_i-1)\ln|V_i|\end{aligned}$$

在 H_0 成立的条件下，$-2\ln\lambda'_k$ 近似服从分布 $\chi_f^2/(1-\boldsymbol{D})$

其中 V_i 为第 i 组样本协方差阵，$V=\frac{1}{n-k}\sum_{i=1}^{k}V_i$，$f=\frac{1}{2}p(p+1)(k-1)$，

$$D=\begin{cases}\dfrac{2p^2+3p-1}{6(p+1)(k-1)}\left(\sum\limits_{i=1}^{k}\dfrac{1}{n_i-1}-\dfrac{1}{n-k}\right),\text{至少有一对 } n_i\neq n_j\\[2ex]\dfrac{(2p^2+3p-1)(k+1)}{6(p+1)(n-k)},\qquad n_1=n_2=\cdots=n_k\end{cases}$$

例 1　某公司欲了解职员的生活状况，随机抽查 15 名员工，测得其日平均面对电脑的小时数（X_1）、日平均运动的小时数（X_2）、日平均睡觉的小时数（X_3），数据如下表。试检验 $H_0:\mu=\mu_0=(7,0.5,7)$　$H_1:\mu\neq\mu_0$。

表 2

序号	X_1	X_2	X_3
1	9	0.5	6
2	7	1	7
3	7	0.2	7
4	8	0.6	8
5	9	0.4	9
6	7.6	1	6
7	5	0.5	7
8	7	0.4	6
9	8	0.5	6
10	7	0.1	6
11	6	1	8
12	4.9	1.2	7
13	5.8	1.4	9
14	6	0.4	6
15	7	0.8	8

根据题意，采用检验统计量

$$\frac{n-p}{(n-1)p}T^2\sim F(p,n-p)$$

其中 $T^2=(n-1)[\sqrt{n}(\overline{X}-\mu_0)'S^{-1}\sqrt{n}(\overline{X}-\mu_0)]$

经计算：

$$n = 15, p = 3, \mu_0 = (7, 0.5, 7), \overline{X} = (6.95, 0.67, 7.07),$$

$$S = \begin{pmatrix} 22.18 & -2.63 & -0.95 \\ -2.63 & 2.01 & 2.43 \\ -0.95 & 2.43 & 16.93 \end{pmatrix} \quad S^{-1} = \begin{pmatrix} 0.0544 & 0.0816 & -0.0087 \\ 0.0816 & 0.7236 & -0.0994 \\ -0.0087 & -0.0994 & 0.0728 \end{pmatrix}$$

可得 $T^2 = 3.5946$，统计的 p 值：0.4152，

在 $\alpha = 0.05$ 的条件下，接受原假设。

例 2 某公司欲了解中层领导与基层职员的生活状况，随机抽查 10 名中层领导(表 3)，15 名基层员工(表 4)，分别测得其日平均面对电脑的小时数(X_1, Y_1)、日平均运动的小时数(X_2, Y_2)、日平均睡觉的小时数(X_3, Y_3)，数据如下表，设两组样本的来自正态总体，分别记为：

$$X_{(\alpha)} \sim N_3(\mu_1, \Sigma) \quad \alpha = 1, \cdots, 10$$
$$Y_{(\alpha)} \sim N_3(\mu_2, \Sigma) \quad \alpha = 1, \cdots, 15$$
$$H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 \neq \mu_2$$

表 3

序号	X_1	X_2	X_3
1	4	0.7	8
2	5	1	7
3	6	0.8	8
4	6	0.9	8
5	9	0.4	9
6	4.5	1	6
7	7	0.7	6
8	6.6	0.7	7.5
9	8	0.5	6
10	6	0.67	7.5

表 4

序号	Y_1	Y_2	Y_3
1	9	0.5	6
2	7	1	7
3	7	0.2	7
4	8	0.6	8
5	9	0.4	9
6	7.6	1	6
7	5	0.5	7
8	7	0.4	6
9	8	0.5	6
10	7	0.1	6

续表

序号	Y_1	Y_2	Y_3
11	6	1	8
12	4.9	1.2	7
13	5.8	1.4	9
14	6	0.4	6
15	7	0.8	8

采用的统计量为：

$$F = \frac{(n+m-2)-p+1}{(n+m-2)p}T^2 \sim F(p, n+m-p+1)$$

其中：$T^2 = (n+m-2)\left[\sqrt{\frac{nm}{n+m}}(\overline{X}-\overline{Y})\right]' S^{-1}\left[\sqrt{\frac{nm}{n+m}}(\overline{X}-\overline{Y})\right]$

$$S = S_1 + S_2$$

$$S_1 = \sum_{\alpha=1}^{n}(X_{(\alpha)}-\overline{X})(X_{(\alpha)}-\overline{X})'$$

$$S_2 = \sum_{\alpha=1}^{m}(Y_{(\alpha)}-\overline{Y})(Y_{(\alpha)}-\overline{Y})'$$

经计算：$n = 10, m = 15, p = 3, \overline{X} = (6.21, 0.737, 7.3), \overline{Y} = (6.95, 0.67, 7.07)$

$$S_1 = \begin{pmatrix} 21.17 & -2.13 & 2.17 \\ -2.13 & 0.35 & -0.53 \\ 2.17 & -0.53 & 9.6 \end{pmatrix}, S_2 = \begin{pmatrix} 22.18 & -2.63 & -0.95 \\ -2.63 & 2.01 & 2.43 \\ -0.95 & 2.43 & 16.93 \end{pmatrix}$$

$$S = S_1 + S_2 = \begin{pmatrix} 43.34 & -4.76 & 1.22 \\ -4.76 & 2.36 & 1.91 \\ 1.22 & 1.91 & 26.53 \end{pmatrix}$$

$$S^{-1} = \begin{pmatrix} 0.0306 & 0.0667 & -0.0062 \\ 0.0667 & 0.5951 & -0.0458 \\ -0.0062 & -0.0458 & 0.0413 \end{pmatrix}$$

$T^2 = 2.173981$，$F = 0.661646$，统计的 p 值：0.58405

在 $\alpha = 0.05$ 的条件下，接受原假设。

本章思考与练习

1. 试列举可运用多元均值检验的实际问题。
2. 试述多元统计分析中的各种均值向量和协差阵检验的基本思想与步骤。
3. 以下是两家上市公司某年的部分收益性及成长性财务指标：

公司简称	深能源 A	深南电 A
净资产收益率 /%	16.85	22
总资产报酬率 /%	12.35	15.3
资产负债率 /%	42.32	46.51
总资产周转率 /%	0.37	0.76
流动资产周转率 /%	1.78	1.77
已获利息倍数 /%	7.18	15.67
销售增长率 /%	45.73	48.11
资本积累率 /%	54.54	19.41

试在显著性水平 $\alpha = 0.05$ 的条件下，检验两家公司的指标向量是否有差异。

第四章 聚类分析

聚类分析(cluster analysis)是研究对样品或指标进行分类的一种多元统计方法。它把分类对象按一定规则分成若干类,这些类非事先给定的,而是根据数据特征确定的。在同一类中这些对象在某种意义上趋向于彼此相似,而在不同类中趋向于不相似。国内也有人称它为群分析、点群分析、簇群分析等。与多元分析的其他方法相比,该方法较为粗糙,理论上还不完善,但应用方面取得了很大成功。通常,人们把聚类分析与回归分析、判别分析称为多元分析的三大方法。

聚类分析给人们提供了丰富多彩的方法进行分类,有系统聚类法、有序样品聚类法、动态聚类法、模糊聚类法、图论聚类法等。本章主要介绍两种常用的聚类分析方法,一种是系统聚类法(Hierarchical Cluster Analysis 也称层次聚类法),另一种是快速聚类分析(K-Means Cluster Analysis 也称 K- 均值聚类法)。其中系统聚类分析根据聚类的对象不同分成两种:一种是对样品(样本)的分类,称为 Q 型,另一种是对变量(指标)的分类,称为 R 型。

4.1 距 离

在介绍聚类方法之前,先引出两样品之间距离的度量。

设有 n 个样品,每个样品测得 p 项指标(变量),资料矩阵为:

$$
\begin{array}{cc}
 & \begin{array}{cccc} Y_1 & Y_2 & \cdots & Y_p \end{array} \\
X = \begin{array}{c} X_1 \\ X_2 \\ \vdots \\ X_n \end{array} & \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}
\end{array}
$$

其中,$x_{ij}(i = 1,\cdots,n;j = 1,\cdots,p)$ 为第 i 个样品的第 j 个指标得观测值;

第 i 个样品 X_i 为矩阵第 i 行所描述,任何两个样品 X_k,X_h 之间的相似性,可以通过矩阵 X 的第 k 行与第 h 行的相似程度来刻画;

第 j 个变量 Y_j 为矩阵的第 j 列所描述,任何两个变量 Y_k,Y_h 之间的相似性,可以通过矩阵 X 的第 k 列与第 h 列的相似程度来刻画。

由于样品分类和指标分类从方法上看基本是一样的,所以两者就不严格分开说明,在这里只介绍两个样品间距离的度量。

4.1.1 聚类数据的标准化处理

在聚类分析中,聚类要素(指标)的选择是十分重要的,它直接影响分类结果的准确性

和可靠性。聚类要素的选择可根据相关的专业的知识及因子分析，主成分分析等统计方法确定。这里，我们假定聚类的要素已经选择。被聚类的对象常常是多个要素构成的，不同要素的数据往往具有不同的单位和量纲，其数值的变异可能是很大的，这就会对分类结果产生影响。因此当分类要素的对象确定之后，在进行聚类分析之前，首先要对聚类要素进行数据处理。

在聚类分析中，常用的聚类要素的数据处理方法有如下几种：

(1) 总和标准化。分别求出各聚类要素所对应的数据的总和，以各要素的数据除以该要素的数据的总和，即

$$x'_{ij}=\frac{x_{ij}}{\sum_{i=1}^{n}x_{ij}}\quad(i=1,\cdots,n;j=1,\cdots,p)$$

(2) 标准差标准化，即

$$x'_{ij}=\frac{x_{ij}-\bar{x}_j}{s_j}\quad(i=1,\cdots,n;j=1,\cdots,p)$$

其中：$\bar{x}_j$，s_j 为第 j 个变量的样本均值，样本标准差。

(3) 极大值标准化，即

$$x'_{ij}=\frac{x_{ij}}{\max_i\{x_{ij}\}}\quad(i=1,\cdots,n;j=1,\cdots,p)$$

(4) 极差标准化，即

$$x'_{ij}=\frac{x_{ij}-\min_i\{x_{ij}\}}{\max_i\{x_{ij}\}-\min_i\{x_{ij}\}}\quad(i=1,\cdots,n;j=1,\cdots,p)$$

4.1.2 样品距离的定义

如果把 n 个样品（X 中的 n 个行）看成 p 维空间中的 n 个点，则第 i 样品 X_i 与第 j 样品 X_j 之间的距离记为 d_{ij}。

常见的距离有：

(1) 绝对值距离：$d_{ij}=\sum_{t=1}^{p}|x_{it}-x_{jt}|$

(2) 欧式距离：$d_{ij}=\sqrt{\sum_{t=1}^{p}(x_{it}-x_{jt})^2}$

(3) 平方欧式距离：$d_{ij}=\sum_{t=1}^{p}(x_{it}-x_{jt})^2$

(4) 切比雪夫距离：$d_{ij}=\mathrm{Max}_t|x_{it}-x_{jt}|$

(5) 明氏(Minkowski) 距离：$d_{ij}=[\sum_{t=1}^{p}|x_{it}-x_{jt}|^q]^{1/q}$

当 $q=1,2$ 时，为绝对值、欧式距离；若趋近无穷时，则为切比雪夫距离。明氏距离在实际的运用很多，但有一些缺点。例如观测值的量纲问题，因此改进得到以下的距离：

(6) 兰氏距离：$d_{ij}(L)=\frac{1}{p}\sum_{t=1}^{p}\frac{|x_{it}-x_{jt}|}{(x_{it}+x_{jt})}$

兰氏距离有助于克服各指标之间量纲的影响，但没有考虑指标的相关性，为了同时克服

这两个缺点，我们引入马氏距离的概念：

(7) 马氏距离：$d_{ij}(M) = (X_i - X_j)'\Sigma^{-1}(X_i - X_j)$

其中：Σ 表示样本的协差阵，即：

$$\Sigma = (\sigma_{ij})_{p\times p}, \quad \sigma_{ij} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{\alpha i} - \bar{x}_i)(x_{\alpha j} - \bar{x}_j) \quad i,j = 1,\cdots,p$$

$$\bar{x}_i = \frac{1}{n}\sum_{\alpha=1}^{n} x_{\alpha i}, \quad \bar{x}_j = \frac{1}{n}\sum_{\alpha=1}^{n} x_{\alpha j}$$

这里 X_i 为样品 X_i 的 p 各指标组成的向量，即原始资料矩阵的第 i 行向量，X_j 类似。

4.2 系统聚类法

正如样品之间的距离可以有不同的定义方法一样，类与类之间的距离也有各种定义。例如可以定义类与类之间的距离为两类之间最近样本的距离，或者定义为两类之间最远样本之间的距离等等。类与类之间用不同的方法定义距离，就产生了不同的系统聚类方法。本节介绍常用的四种方法，即最短距离法、最长距离法、重心法、类平均法。系统聚类分析尽管方法很多，但归类的步骤基本上是一样的，所不同的仅是类与类之间距离的不同定义方法。

系统聚类法的基本过程：首先将 n 个样品看成 n 类（一个类包含一个样品），然后将性质最接近的两类合并成一个新类，我们得到 $n-1$ 类，再从中找出最接近的两类加以合并成了 $n-2$ 类，如此下去，最后所有的样品均在一类，将上述并类过程画成一张图（称为聚类图），根据实际的研究目的，便可决定分多少类，每类各有什么样品。

以下用 d_{ij} 表示样品 X_i 与 X_j 之间的距离，用 $D(p,q)$ 表示类 G_p 和 G_q 的距离。

4.2.1 类间的距离

下边是一些类与类之间距离的定义：

(1) 最短距离法

定义距离：$D(p,q) = \text{Min}\{d_{ij}: X_i \in G_p, X_j \in G_q\}$

等于 G_p 和 G_q 最为邻近的两个样品之间的距离。

(2) 最长距离法

定义距离：$D(p,q) = \text{Max}\{d_{ij}: X_i \in G_p, X_j \in G_q\}$

等于 G_p 和 G_q 最远的两个样品之间的距离

(3) 重心法

定义距离平方：$D_c^2(p,q) = d^2_{\bar{X}_p \bar{X}_q}$

等于 G_p 和 G_q 两个重心之间的距离，这里，$\bar{X}_p = \sum_{X\in G_p} X/n, \bar{X}_q = \sum_{X\in G_q} X/m, n, m$ 分别 G_p 和 G_q 的元素的个数。

(4) 类平均法

定义距离平方：$D^2(p,q) = \frac{1}{n_p n_q}\sum_{X_i\in G_p}\sum_{X_j\in G_q} d_{ij}^2$

等于 G_p 和 G_q 中任意两个样品距离的平均，其中 n_p, n_q 为 G_p, G_q 的样品数。

4.2.2 四种系统聚类法

系统聚类分析尽管方法很多，但归类的步骤基本是一样，所不同的仅是类与类之间的距离有不同的定义方法。

(一) 最短距离法

最短距离法类与类之间的距离：$D(p,q) = \text{Min}\{d_{ij}: X_i \in G_p, X_j \in G_q\}$

设类 G_p 与 G_q 合并成一个新类记为 G_r，则任一类 G_k 与 G_r 的距离：

$$\begin{aligned} D(k,r) &= \text{Min}\{d_{ij}: X_i \in G_k, X_j \in G_r\} \\ &= \text{Min}\{\text{Min}\{d_{ij}: X_i \in G_k, X_j \in G_p\}, \text{Min}\{d_{ij}: X_i \in G_k, X_j \in G_q\}\} \\ &= \text{Min}\{D(k,p), D(k,q)\} \end{aligned}$$

最短距离法聚类的步骤如下：

(1) 定义样品之间的距离，计算样品两两距离，得一距离阵记为 $D_{(0)}$，开始每个样品自成一类，显然这时 $D_{ij} = d_{ij}$

(2) 找出 $D_{(0)}$ 的非对角线最小元素，设为 D_{pq}，则将 G_p 和 G_q 合并成一个新类，记为 G_r，则 $G_r = \{G_p, G_q\}$

(3) 给出计算新类与其他类的距离公式

$$D_{kr} = \text{Min}\{D_{kp}, D_{kq}\}$$

将 $D_{(0)}$ 中第 p,q 行及第 p,q 列用上面的公式并成一个新行新列，新行新列对应 G_r，所得的矩阵记为 $D_{(1)}$

(4) 对 $D_{(1)}$ 重复上述对 $D_{(0)}$ 的(2)(3)两步得到 $D_{(2)}$；如此下去，直到所有的元素并成一类为止。

下边通过一个例子说明各种聚类法。

例 1 为了研究黑龙江、内蒙古、江苏，广东，广西 5 个省份 2010 年第三产业的分布情况，根据调查资料做类型划分。指标名称及原始数据见表 1。资料来源中国统计年鉴(2011年)。

X_1：交通运输、仓储和邮政业

X_2：批发和零售业

X_3：住宿和餐饮业

X_4：金融业

X_5：房地产业

X_6：其他(第三产业)

表 1 2010 年 5 个省份第三产业的生产总值 (亿元)

	X_1	X_2	X_3	X_4	X_5	X_6
黑龙江	469.31	880.83	240.13	288.19	370.79	1612.3
内蒙古	875.61	1052	332.24	346.44	309.25	1293.5
江苏	1768.3	4447.5	710.98	2105.9	2601	5497.8
广东	1825.29	4647.8	1074.9	2658.8	2814	7690.9
广西	480.17	656.83	241.34	384.53	405.79	1214.5

(1) 采用极差标准化数据：

	X_1	X_2	X_3	X_4	X_5	X_6
黑龙江	1	0.9439	1	1	0.9754	0.9386
内蒙古	0.70036	0.901	0.8897	0.9754	1	0.9878
江苏	0.04203	0.0502	0.4359	0.2332	0.085	0.3386
广东	0	0	0	0	0	0
广西	0.99199	1	0.9986	0.9594	0.9615	1

(2) 再计算 5 个省份之间的平方欧氏距离，用 D_0 表示 5 个省份两两的距离矩阵：

$$D_0 = \begin{bmatrix} 0 & 0.1074 & 3.7753 & 5.7233 & 0.0088 \\ & 0 & 3.1726 & 5.021 & 0.1086 \\ & & 0 & 0.3706 & 3.854 \\ & & & 0 & 5.8259 \\ & & & & 0 \end{bmatrix}$$

距离矩阵 $D_0 = (d_{ij})_{5\times5}$，$D_0$ 是对称矩阵：$d_{ij} = d_{ji}$，其中：d_{ij} 代表第 i 个省份与第 j 个省份数据的欧氏距离，例如：d_{12} 代表黑龙江样品与内蒙古样品的欧氏距离。

(3) 开始聚类

① 初始化

开始 5 类：$G_1 = \{$黑龙江 1$\}$、$G_2 = \{$内蒙古 2$\}$、$G_3 = \{$江苏 3$\}$、$G_4 = \{$广东 4$\}$、$G_5 = \{$广西 5$\}$，由类间最短距离法的定义，这时：

$$D_0(i,j) = d_{ij} \quad i,j = 1,\cdots,5$$

其中：$D_0(i,j)$ 表示 G_i，G_j 的类间距离。

② 合并类

D_0 中非对角线的最小的元素是 $D_0(1,5) = 0.0088$，故将类 G_1 和 G_5 合并成一新类 $G_6 = \{1,5\}$，接下来继续计算 G_6 与 G_2，G_3，G_4 之间的距离。

③ 计算新类距离矩阵

利用

$$D_1(6,j) = \text{Min}\{D_0(1,j), D_0(5,j)\} \quad j = 2,3,4$$

得到的新距离矩阵

$$D_1 = \begin{bmatrix} & G_6 & G_2 & G_3 & G_4 \\ G_6 & 0 & 0.1074 & 3.7753 & 5.7233 \\ G_2 & & 0 & 3.1726 & 5.021 \\ G_3 & & & 0 & 0.3706 \\ G_4 & & & & 0 \end{bmatrix}$$

在上表中，找出非对角线的类间最小距离：$D_1(G_6, G_2) = 0.1074$，合并类 $G_6 = \{1,5\}$ 与类 $G_2 = \{2\}$ 得到新类：$G_7 = \{1,5,2\}$，再利用类间最小距离法公式：

$$D_2(7,j) = \text{Min}\{D_1(2,j), D_1(6,j)\} \quad j = 3,4$$

$$D_2 = \begin{bmatrix} & G_7 & G_3 & G_4 \\ G_7 & 0 & 3.1726 & 5.021 \\ G_3 & & 0 & 0.3706 \\ G_4 & & & 0 \end{bmatrix}$$

类间的最小距离：$D_2(G_3,G_4)=0.3706$，合并类 G_3，G_4 得到新类：$G_8=\{3,4\}$。

此时，我们有两个不同的类：$G_7=\{1,5,2\}$，$G_8=\{3,4\}$，两者合并成一个大类的聚类系统。

最后，根据计算的过程，可得到谱系聚类图如下：



图 1

从图 1 可看出：若分成二类：{黑龙江，广西，内蒙古}{江苏，广东}；若分成三类：{黑龙江，广西}，{内蒙古}，{江苏，广东}。

聚类分析是一种探索性方法，确定分类数的问题是迄今为止未完全解决的问题之一。实际应用中，主要根据研究的目的，从实用的角度出发，选择合适的分类。如本例若研究的对象是第三产业发达与第三产业不发达地区的研究，那么可分成二类；若研究的对象是：发达，欠发达，落后，则可分成三类。

（二）最长距离法

最长距离法类与类之间的距离：$D(p,q)=Max\{d_{ij}:X_i\in G_p,X_j\in G_q\}$。

最长距离法与最短距离法的并类步骤完全一样，也是将各样品先自成一类，然后将非对角线上最小元素对应的两类合并，直至所有的样品全归为一类为止。所不同的是类与类之间的距离定义不同。设某一步将 G_p 于 G_q 合并为 G_r，则任一类 G_k 与 G_r 的距离用最长距离公式为：

$$\begin{aligned}D(k,r)&=\mathrm{Max}\{d_{ij}:X_i\in G_k,X_j\in G_r\}\\&=\mathrm{Max}\{\mathrm{Max}\{d_{ij}:X_i\in G_k,X_j\in G_p\},\mathrm{Max}\{d_{ij}:X_i\in G_k,X_j\in G_q\}\}\\&=\mathrm{Max}\{D(k,p),D(k,q)\}\end{aligned}$$

将例 1 应用最长距离法如下：

① $$D_0=\begin{bmatrix}0 & 0.1074 & 3.7753 & 5.7233 & 0.0088\\ & 0 & 3.1726 & 5.021 & 0.1086\\ & & 0 & 0.3706 & 3.854\\ & & & 0 & 5.8259\\ & & & & 0\end{bmatrix}$$

D_0 中非对角线的最小的元素是 $D_0(1,5)=0.0088$，故将类 G_1 和 G_5 合并成一新类 $G_6=\{1,5\}$，接下来按最长距离法计算 G_6 与 G_2，G_3，G_4 之间的距离。

② $$D_1=\begin{bmatrix} & G_6 & G_2 & G_3 & G_4\\ G_6 & 0 & 0.1086 & 3.854 & 5.8259\\ G_2 & & 0 & 3.1726 & 5.021\\ G_3 & & & 0 & 0.3706\\ G_4 & & & & 0\end{bmatrix}$$

找出非对角线的类间最小距离：$D_1(G_6, G_2) = 0.1086$，合并类 $G_6 = \{1,5\}$ 与类 $G_2 = \{2\}$ 得到新类：$G_7 = \{1,5,2\}$，

③ $$D_2 = \begin{bmatrix} & G_7 & G_3 & G_4 \\ G_7 & 0 & 3.854 & 5.8259 \\ G_3 & & 0 & 0.3706 \\ G_4 & & & 0 \end{bmatrix}$$

类间的最小距离：$D_2(G_3, G_4) = 0.3706$，合并类 G_3，G_4 得到新类：$G_8 = \{3,4\}$。

此时，我们有两个不同的类：$G_7 = \{1,5,2\}$，$G_8 = \{3,4\}$，两者合并成一个大类的聚类系统。其聚类图如下，与最短距离法分类情况一致，只是并类的距离不同。

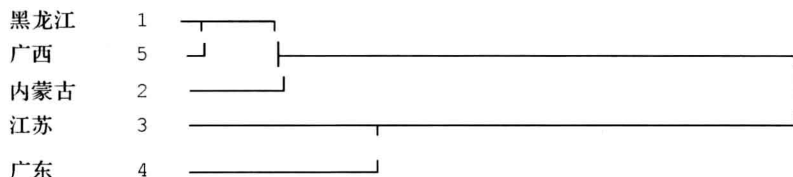


图 2

(3) 重心法

重心法类与类之间的距离平方：$D_c^2(p,q) = d^2_{\bar{X}_p \bar{X}_q}$，重心法定义两类之间的距离就是两类重心之间的距离。

将例 1 应用重心法如下：

① $$D_0 = \begin{bmatrix} 0 & 0.1074 & 3.7753 & 5.7233 & 0.0088 \\ & 0 & 3.1726 & 5.021 & 0.1086 \\ & & 0 & 0.3706 & 3.854 \\ & & & 0 & 5.8259 \\ & & & & 0 \end{bmatrix}$$

D_0 中非对角线的最小的元素是 $D_0(1,5) = 0.0088$，故将类 G_1 和 G_5 合并成一新类 $G_6 = \{1,5\}$，其重心：$\bar{x}_6 = \{0.995996, 0.97194, 0.99928, 0.97968, 0.96844, 0.96928\}$，接下来按重心法计算 G_6 与 G_2，G_3，G_4 之间的距离。

② $$D_1 = \begin{bmatrix} & G_6 & G_2 & G_3 & G_4 \\ G_6 & 0 & 0.1058 & 3.8124 & 5.7724 \\ G_2 & & 0 & 3.1726 & 5.021 \\ G_3 & & & 0 & 0.3706 \\ G_4 & & & & 0 \end{bmatrix}$$

找出非对角线的类间最小距离：$D_1(G_6, G_2) = 0.1058$，合并类 $G_6 = \{1,5\}$ 与类 $G_2 = \{2\}$ 得到新类：$G_7 = \{1,5,2\}$，计算其重心：

$\bar{x}_7 = \{0.89745, 0.94829, 0.96273, 0.97826, 0.96896, 0.97545\}$

③
$$D_2 = \begin{bmatrix} & G_7 & G_3 & G_4 \\ G_7 & 0 & 3.5756 & 5.4984 \\ G_3 & & 0 & 0.3706 \\ G_4 & & & 0 \end{bmatrix}$$

类间的最小距离：$D_2(G_3,G_4)=0.3706$，合并类 G_3，G_4 得到新类：$G_8=\{3,4\}$。

此时，我们有两个不同的类：$G_7=\{1,5,2\}$，$G_8=\{3,4\}$，两者合并成一个大类的聚类系统。其聚类图如下，与前边两个分类情况一致，只是并类的距离不同。

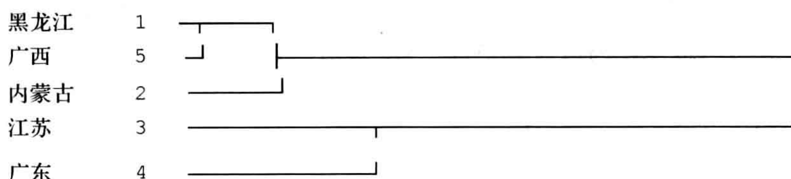


图 3

(4) 类平均法

类平均法类与类之间的距离平方：$D^2(p,q)=\frac{1}{n_p n_q}\sum_{X_i\in G_p}\sum_{X_j\in G_q} d_{ij}^2$，它定义两类之间的距离平方为这两类元素两两之间距离平方的平均。

将例 1 应用类平均法如下：

①
$$D_0 = \begin{bmatrix} 0 & 0.1074 & 3.7753 & 5.7233 & 0.0088 \\ & 0 & 3.1726 & 5.021 & 0.1086 \\ & & 0 & 0.3706 & 3.854 \\ & & & 0 & 5.8259 \\ & & & & 0 \end{bmatrix}$$

D_0 中非对角线的最小的元素是 $D_0(1,5)=0.0088$，故将类 G_1 和 G_5 合并成一新类 $G_6=\{1,5\}$，接下来按类平均法计算 G_6 与 G_2，G_3，G_4 之间的距离。

②
$$D_1 = \begin{bmatrix} & G_6 & G_2 & G_3 & G_4 \\ G_6 & 0 & 0.108 & 3.8148 & 5.7738 \\ G_2 & & 0 & 3.1726 & 5.021 \\ G_3 & & & 0 & 0.3706 \\ G_4 & & & & 0 \end{bmatrix}$$

找出非对角线的类间最小距离：$D_1(G_6,G_2)=0.108$，合并类 $G_6=\{1,5\}$ 与类 $G_2=\{2\}$ 得到新类：$G_7=\{1,5,2\}$，

③
$$D_2 = \begin{bmatrix} & G_7 & G_3 & G_4 \\ G_7 & 0 & 3.6135 & 5.535 \\ G_3 & & 0 & 0.3706 \\ G_4 & & & 0 \end{bmatrix}$$

类间的最小距离：$D_2(G_3,G_4)=0.3706$，合并类 G_3，G_4 得到新类：$G_8=\{3,4\}$。

此时，我们有两个不同的类：$G_7=\{1,5,2\}$，$G_8=\{3,4\}$，两者合并成一个大类的聚类系

统。其聚类图如下。

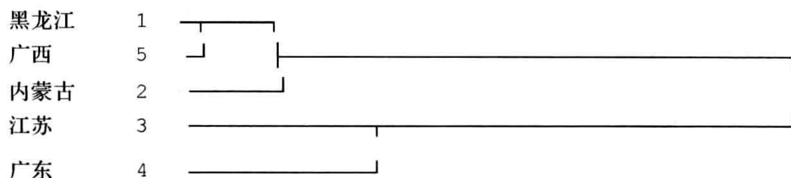


图 4

聚类分析所使用方法的不同，常常会得到不同的结论。不同研究者对于同一组数据进行聚类分析，所得到的聚类数未必一致。因此我们说聚类分析是一种探索性的分析方法。

4.3 K-均值聚类法

K-均值聚类法是一种非谱系聚类法，它是把样品聚集成 k 个类的集合。类的个数 k 可以预先给定或者在聚类过程中确定。该方法可用于比系统聚类法大得多的数据组。如果观察值的个数多或文件非常庞大(通常观察值在 200 个以上)，则宜采用 K-均值聚类法。

K-均值聚类法算法的工作原理：算法首先随机从数据集中选取 K 个点作为初始聚类中心，然后计算各个样本到聚类中的距离，把样本归到离它最近的那个聚类中心所在的类。计算新形成的每一个聚类的数据对象的平均值来得到新的聚类中心，如果相邻两次的聚类中心没有任何变化，说明样本调整结束，聚类准则函数已经收敛。本算法的一个特点是在每次迭代中都要考察每个样本的分类是否正确。若不正确，就要调整，在全部样本调整完后，再修改聚类中心，进入下一次迭代。如果在一次迭代算法中，所有的样本被正确分类，则不会有调整，聚类中心也不会有任何变化，这标志着已经收敛，因此算法结束。

K-means 聚类算法的一般步骤：

(1) 从 n 个数据对象任意选择 k 个对象作为初始聚类中心；

(2) 根据每个聚类对象的均值(中心对象)，计算每个对象与这些中心对象的距离；并根据最小距离重新对相应对象进行划分；

(3) 重新计算每个(有变化) 聚类的均值(中心对象)；

(4) 循环(2) 到(3) 直到每个聚类不再发生变化为止。

样品的最终聚类在某种程度上依赖于最初的划分，或种子点的选择，为了检验聚类的稳定性，可用一个新的初始分类重新检验整个聚类算法。如果最终分类与原来一样，则不必再行计算；否则，须另行考虑聚类算法。

本章思考与练习

1. 试述系统聚类的基本思想。

2. 在进行系统聚类时，不同的类间距离计算方法有何区别。

3. 试述系统聚类法和 K 均值法的异同。

4. 下表给出了某农业生态经济系统各个区域单元的有关数据，试运用系统聚类法，对该农业生态经济系统进行聚类分析：

样本序号	人口密度 x_1/(人·km^{-2})	人均耕地面积 x_2/hm^2	森林覆盖率 x_3/%	农民人均纯收入 x_4/(元·人$^{-1}$)	人均粮食产量 x_5/(kg·人$^{-1}$)	经济作物占农作物播面比例 x_6/%	耕地占土地面积比率 x_7/%	果园与林地面积之比 x_8/%	灌溉田占耕地面积之比 x_9/%
1	363.912	0.352	16.101	192.11	295.34	26.724	18.492	2.231	26.262
2	141.503	1.684	24.301	1 752.35	452.26	32.314	14.464	1.455	27.066
3	100.695	1.067	65.601	1 181.54	270.12	18.266	0.162	7.474	12.489
4	143.739	1.336	33.205	1 436.12	354.26	17.486	11.805	1.892	17.534
5	131.412	1.623	16.607	1 405.09	586.59	40.683	14.401	0.303	22.932
6	68.337	2.032	76.204	1 540.29	216.39	8.128	4.065	0.011	4.861
7	95.416	0.801	71.106	926.35	291.52	8.135	4.063	0.012	4.862
8	62.901	1.652	73.307	1 501.24	225.25	18.352	2.645	0.034	3.201
9	86.624	0.841	68.904	897.36	196.37	16.861	5.176	0.055	6.167
10	91.394	0.812	66.502	911.24	226.51	18.279	5.643	0.076	4.477
11	76.912	0.858	50.302	103.52	217.09	19.793	4.881	0.001	6.165
12	51.274	1.041	64.609	968.33	181.38	4.005	4.066	0.015	5.402
13	68.831	0.836	62.804	957.14	194.04	9.11	4.484	0.002	5.79
14	77.301	0.623	60.102	824.37	188.09	19.409	5.721	5.055	8.413
15	76.948	1.022	68.001	1255.42	211.55	11.102	3.133	0.01	3.425
16	99.265	0.654	60.702	1251.03	220.91	4.383	4.615	0.011	5.593
17	118.505	0.661	63.304	1246.47	242.16	10.706	6.053	0.154	8.701
18	141.473	0.737	54.206	814.21	193.46	11.419	6.442	0.012	12.945
19	137.761	0.598	55.901	1124.05	228.44	9.521	7.881	0.069	12.654
20	117.612	1.245	54.503	805.67	175.23	18.106	5.789	0.048	8.461
21	122.781	0.731	49.102	1313.11	236.29	26.724	7.162	0.092	10.078

第五章　判别分析

5.1　判别分析简介

判别分析(discriminant analysis) 又称“分辨法”，根据已知类别的样本所能提供的信息，总结出分类的规律性，建立判别公式和判别准则，判别新的样本点所属类型，是判别个体所属群体的一种统计方法。

聚类分析与判别分析都是多元统计中研究事物分类的基本方法，两者有何区别?

主要不同点就是，在聚类分析中一般人们事先并不知道或一定要明确应该分成几类，全根据数据来确定。而在判别分析中，至少有一个已经明确知道类别的“训练样本”，利用这些数据集，就可以建立判别准则，并通过预测变量来对未知类别的观测值进行判别。

判别分析的基本思路：设有总体 G_1、G_2、…、G_k，根据样本 → 建立判别法则(判别函数) → 判别新的样品属于哪一个总体。判别新样本所属类别分类数目一般根据研究目的而定。

判别函数的一般形式为

$$Y = a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

其中，Y 为判别分数(判别值)；

$x_1, x_2 \cdots x_n$ 为反映研究对象特征的变量；

$a_1, a_2 \cdots a_n$ 为各变量的系数，称为判别系数

判别分析内容很丰富，方法很多。判断分析按判别的总体数来区分，有两个总体判别分析和多总体判别分析；按区分不同总体所用的数学模型来分，有线性判别和非线性判别；按判别时所处理的变量方法不同，有逐步判别和序贯判别等。判别分析可以从不同角度提出问题，因此有不同的判别准则，如马氏距离最小准则、*Fisher* 准则、平均损失最小准则、最小平方准则、最大似然准则、最大概率准则等等，按判别准则的不同又提出多种判别方法。

本章仅介绍常用的几种判别分析方法：距离判别法、*Fisher* 判别法、*Bayes* 判别法和逐步判别法。

5.2　距离判别法

距离判别的基本思想：首先根据已知分类的数据，分别计算各类的重心，即分组(类) 均值，距离判别准则是对于任给一次观测值，若它与第 i 类的重心距离最近，就认为它来自第 i 类。因此，距离判别法又称为最邻近方法(*nearest neighbor method*)。距离判别法对各类总体

的分布没有特定的要求，适用于任意分布的数据资料。

5.2.1 两组距离判别

设有两组总体 G_1 和 G_2，相应两个样品容量为 n_1, n_2 的样品，从两个总体分别取得 p 维观察值，

$$\{G_1 : X_1^{(1)} \cdots X_1^{(1)}\}.$$

$$\overline{X}_i^{(1)} = \sum_{t=1}^{n_1} x_{n} / n_1 \quad i = 1, \cdots, p$$

$$\overline{X}^{(1)} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i^{(1)}$$

$$\overline{X}^{(2)} = \frac{1}{n_2} \sum_{i=1}^{n_2} X_i^{(2)}$$

$$\overline{X}_i^{(2)} = \sum_{i=1}^{n_2} X_t i / n_2, i = 1, \cdots, p$$

每个样品观测 p 个指标得观测数据如下：

则总体 G_1 的样本数据为：

$$\begin{matrix} X_1^{(1)} \\ X_2^{(1)} \\ \vdots \\ X_{n_1}^{(1)} \end{matrix} \begin{bmatrix} x_{11}^{(1)} & x_{12}^{(1)} & \cdots & x_{1p}^{(1)} \\ x_{21}^{(1)} & x_{22}^{(1)} & \cdots & x_{2p}^{(1)} \\ \cdots & \cdots & & \cdots \\ x_{n_1 1}^{(1)} & x_{n_1 2}^{(1)} & \cdots & x_{n_1 p}^{(1)} \end{bmatrix}$$

记：

$$\overline{X}^{(1)} = (\overline{x}_1^{(1)}, \overline{x}_2^{(1)} \cdots \overline{x}_p^{(1)})'$$

总体 G_2 的样本数据为：

$$\begin{matrix} X_1^{(2)} \\ X_2^{(2)} \\ \vdots \\ X_{n_2}^{(2)} \end{matrix} \begin{bmatrix} x_{11}^{(2)} & x_{12}^{(2)} & \cdots & x_{1p}^{(2)} \\ x_{21}^{(2)} & x_{22}^{(2)} & \cdots & x_{2p}^{(2)} \\ \cdots & \cdots & & \cdots \\ x_{n_2 1}^{(2)} & x_{n_2 2}^{(2)} & \cdots & x_{n_2 p}^{(2)} \end{bmatrix}$$

记：

$$\overline{X}^{(2)} = (\overline{x}_1^{(2)}, \overline{x}_2^{(2)}, \cdots, \overline{x}_p^{(2)})'$$

其中：$X_i^{(1)}$，$X_i^{(2)}$ 列向量，例如：$X_1^{(1)} = (x_{11}^{(1)}, x_{12}^{(1)}, \cdots, x_{1p}^{(1)})'$，本章随后章节皆如此，不再说明。

现任取一个新个体，观察值 X 为 $X = (x_1, x_2, \cdots, x_p)'$，问 X 应判归于哪一类？

首先计算 X 到与 G_1、G_2 总体的距离，分别记为 $D(X, G_1)$、$D(X, G_2)$，按距离最近准则判别归类，则可写成：

$$\begin{cases} X \in G_1 & D(X, G_1) < D(X, G_2) \\ X \in G_2 & D(X, G_1) > D(X, G_2) \\ X \text{ 待判} & D(X, G_1) = D(X, G_2) \end{cases}$$

如果距离定义采用欧氏距离，则可计算出

$$D(X, G_1) = \sqrt{(X - \overline{X}^{(1)})'(X - \overline{X}^{(1)})}$$

$$D(X, G_2) = \sqrt{(X - \overline{X}^{(2)})'(X - \overline{X}^{(2)})}$$

然后比较 $D(X,G_1)$ 和 $D(X,G_2)$ 大小，按距离最近准则判别归类。

实际应用中，考虑到判别分析常涉及多个变量，且变量之间可能相关，故多用马氏距离。马氏距离公式为：

$$D^2(X,G_1)=(X-\mu^{(1)})'(\sum^{(1)})^{-1}(X-\mu^{(1)})$$

$$D^2(X,G_2)=(X-\mu^{(2)})'(\sum^{(2)})^{-1}(X-\mu^{(2)})$$

其中 $\mu^{(1)}$、$\mu^{(2)}$、$\sum^{(1)}$、$\sum^{(2)}$ 分别是 G_1 和 G_2 的均值向量和协差阵。

这时的判别准则分两种情况给出：

(1) 当 $\sum^{(1)}=\sum^{(2)}=\sum$ 时

$$\begin{aligned}&D^2(X,G_2)-D^2(X,G_1)\\&=(X-\mu^{(2)})'(\sum)^{-1}(X-\mu^{(2)})-(X-\mu^{(1)})'(\sum)^{-1}(X-\mu^{(1)})\\&=2\left[X-\frac{1}{2}(\mu^{(1)}+\mu^{(2)})\right]'\sum^{-1}(\mu^{(1)}-\mu^{(2)})\end{aligned}$$

记 $\bar{\mu}=\frac{1}{2}(\mu^{(1)}+\mu^{(2)})$，$\alpha=\sum^{-1}(\mu^{(1)}-\mu^{(2)})$

记
$$\begin{aligned}W(X)&=(X-\bar{\mu})'\sum^{-1}(\bar{\mu}^{(1)}-\bar{\mu}^{(2)})=(X-\bar{\mu})'\alpha\\&=a_1(x_1-\bar{\mu}^{(1)})+\cdots+a_p(x_p-\bar{\mu}^{(p)})\end{aligned}$$

于是判别准则写成：

$$\begin{cases}X\in G_1 & W(X)>0\\X\in G_2 & W(X)<0\\X\text{ 待判} & W(X)=0\end{cases}$$

该规则取决于 $W(X)$ 的值，因此 $W(X)$ 被称为判别函数，由于它是线性函数，又称为线性判别函数，α 称为判别系数(类似于回归系数)。

当 $\mu^{(1)}$，$\mu^{(2)}$，$\sum$ 未知时，可通过样本来估计。

设 $X_1^{(i)},X_2^{(i)},\cdots,X_{n_i}^{(i)}$ 来自 G_i 的样本，$i=1,2$。

$$\hat{\mu}^{(1)}=\frac{1}{n_1}\sum_{i=1}^{n_1}X_i^{(1)}=\bar{X}^{(1)},\hat{\mu}^{(2)}=\frac{1}{n_2}\sum_{i=1}^{n_2}X_i^{(2)}=\bar{X}^{(2)},\hat{\sum}=\frac{1}{n_1+n_2-2}(S_1+S_2)$$

其中：$S_i=\sum_{t=1}^{n_i}(X_t^{(i)}-\bar{X}^{(i)})(X_t^{(i)}-\bar{X}^{(i)})'$，$\bar{X}=\frac{1}{2}(\bar{X}^{(1)}+\bar{X}^{(2)})$

线性判别函数为：

$$W(X)=(X-\bar{X})'\hat{\sum}^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)})$$

(2) 当 $\sum^{(1)}\neq\sum^{(2)}$ 时

按照距离最近准则，类似地有：

$$\begin{cases}X\in G_1 & D^2(X,G_1)<D^2(X,G_2)\\X\in G_2 & D^2(X,G_1)>D^2(X,G_2)\\X\text{ 待判} & D^2(X,G_1)=D^2(X,G_2)\end{cases}$$

仍然用 $W(X)=D^2(X,G_2)-D^2(X,G_1)$

$$= (X-G^{(2)})'(\sum^{(2)})^{-1}(X-G^{(2)}) - (X-G^{(1)})'(\sum^{(1)})^{-1}(X-G^{(1)}).$$

$\sum^{(1)} \sum^{(2)}$ 可用估计量$\frac{S_1}{n_1-1}$,$\frac{S_2}{n_2-1}$来代替.

作为判别函数,此时的判别函数是 X 的二次函数。

(3) 检验

由于判别分析是假设两组样品是取自不同总体,如果两个总体的均值向量在统计上差异不显著,则进行判别分析意义不大。所以,两组判别分析的检验,实际就是要检验验两个正态总体的均值向量是否相等,相关内容可参考第三章。

5.2.2 多个总体的距离判别法

设有 k 个总体 $G_1 \cdots G_k$,它们的均值和协差阵分别为:$\mu^{(i)}$, $\sum^{(i)}$, $i = 1,\cdots,k$,

相应抽出样品个数为 $n_1 \cdots n_k (n_1 + \cdots + n_k = n)$,每个样品观测 p 个指标得观测数据如下,

总体 G_1 的样本数据为:

$$\begin{matrix} X_1^{(1)} \\ X_2^{(1)} \\ \vdots \\ X_{n_1}^{(1)} \end{matrix} \begin{bmatrix} x_{11}^{(1)} & x_{12}^{(1)} & \cdots & x_{1p}^{(1)} \\ x_{21}^{(1)} & x_{22}^{(1)} & \cdots & x_{2p}^{(1)} \\ \cdots & \cdots & & \cdots \\ x_{n_1 1}^{(1)} & x_{n_1 2}^{(1)} & \cdots & x_{n_1 p}^{(1)} \end{bmatrix}$$

总体 G_k 的样本数据为:

$$\begin{matrix} X_1^{(k)} \\ X_2^{(k)} \\ \vdots \\ X_{n_k}^{(k)} \end{matrix} \begin{bmatrix} x_{11}^{(k)} & x_{12}^{(k)} & \cdots & x_{1p}^{(k)} \\ x_{21}^{(k)} & x_{22}^{(k)} & \cdots & x_{2p}^{(k)} \\ \cdots & \cdots & & \cdots \\ x_{n_k 1}^{(k)} & x_{n_k 2}^{(k)} & \cdots & x_{n_k p}^{(k)} \end{bmatrix}$$

记总体的样本指标平均值为:

$$\overline{X}^{(i)} = (\overline{x}_1^{(i)}, \overline{x}_2^{(i)}, \cdots, \overline{x}_p^{(i)})', i = 1,2,\cdots,k。$$

任取一个样品,测得其指标值 $X = (x_1, \cdots, x_p)'$,问 X 应判归哪一类?

(1) 当$\sum^{(1)} = \cdots = \sum^{(k)} = \sum$ 时

此时 $D^2(X,G_i) = (X-\mu^{(i)})' \sum^{-1} (X-\mu^{(i)}), i = 1,2,\cdots,k$

判别函数为:

$$W_{ij}(X) = \frac{1}{2}[D^2(X,G_j) - D^2(X,G_i)], i,j = 1,2,\cdots,k$$

相应的判别准则为:

$$\begin{cases} X \in G_i, & \text{当 } W_{ij}(X) > 0 \text{ 时,对于一切 } j \neq i \\ \text{待判}, & \text{若有一个 } W_{ij}(X) = 0 \end{cases}$$

当 $\mu^{(1)}, \cdots, \mu^{(k)}, \sum$ 未知时,可通过样本来估计。

设 $X_1^{(i)}, X_2^{(i)}, \cdots, X_{n_i}^{(i)}$ 来自 G_i 的样本,$i = 1,\cdots,k$。

$$\hat{\mu}_i = \frac{1}{n_i}\sum_{a=1}^{n_i} X_a^{(i)} = \overline{X}^{(i)}, i = 1,\cdots,k, \quad \hat{\Sigma} = \frac{1}{n-k}\sum_{i=1}^{k} S_i$$

其中：$S_i = \sum_{t=1}^{n_i}(X_t^{(i)} - \overline{X}^{(i)})(X_t^{(i)} - \overline{X}^{(i)})'$

(2) 当$\sum^{(1)},\cdots,\sum^{(k)}$不相等时

此时判别函数为

$$W_{ij}(X) = (X - \overline{X}^{(j)})'(\textstyle\sum^{(j)})^{-1}(X - \overline{X}^{(j)}) - (X - \overline{X}^{(i)})'(\textstyle\sum^{(i)})^{-1}(X - \overline{X}^{(i)})$$

相应的判别准则为：

$$X \in G_i, \quad 当 W_{ij}(X) > 0 时，对于一切 j \neq i$$

待判，若有一个$W_{ij}(X) = 0$其中$\sum^{(i)}$可用它的估计量$S_i/n_i - 1$来代替。

例 1 为研究 2010 年中国各省年生产总值状况，今选择经济相对发达、经济中等发展水平的省份各五个作为两组样品，另选取三个省份作为待判样品做距离判别分析，判别指标及原始数据见表 1-3。资料来源中国统计年鉴(2011 年)。

x_1：第一产业生产总值

x_2：第二产业生产总值

x_3：第三产业生产总值

表 1 （单位：亿元）

G_1	第一产业	第二产业	第三产业
河北	2562.81	10707.68	7123.77
山西	554.48	5234	3412.38
辽宁	1631.08	9976.82	6849.37
内蒙古	1095.28	6367.69	4209.02
黑龙江	1302.9	5204.11	3861.59

表 2 （单位：亿元）

G_2	第一产业	第二产业	第三产业
上海	114.15	7218.32	9833.51
江苏	2540.1	21753.93	17131.45
浙江	1360.56	14297.93	12063.82
山东	3588.28	21238.49	14343.14
广东	2286.98	23014.53	20711.55

表 3 （单位：亿元）

待判样品	第一产业	第二产业	第三产业
安徽	1729.02	6436.62	4193.68
云南	1108.38	3223.49	2892.31
福建	1363.67	7522.83	5850.62

本例中变量个数$p = 3$，两类总体各有 5 个样品，$n_1 = 5$，$n_2 = 5$，有 3 个待判样品，为方便起见，假定两个总体协差阵相等。

两组线性判别的计算过程如下：

(1) $\overline{X}^{(1)} = \begin{pmatrix} 1429.31 \\ 7498.06 \\ 5091.226 \end{pmatrix}$ $\quad\overline{X}^{(2)} = \begin{pmatrix} 1978.01 \\ 17504.64 \\ 14816.69 \end{pmatrix}$

(2) 计算样本协差阵，求出 $\hat{\sum}$

$$S_1 = \sum_{t=1}^{n_1}(X_t^{(1)} - \overline{X}^{(1)})(X_t^{(1)} - \overline{X}^{(1)})'$$

$$= \begin{pmatrix} 2218416 & 6786467 & 4577456 \\ 6786467 & 28111822 & 18500662 \\ 4577456 & 18500662 & 12331121 \end{pmatrix}$$

$$S_2 = \sum_{t=1}^{n_2}(X_t^{(2)} - \overline{X}^{(2)})(X_t^{(2)} - \overline{X}^{(2)})'$$

$$= \begin{pmatrix} 685996 & 31255624 & 13347604 \\ 31255624 & 178448357.3 & 100634191 \\ 13347604 & 100634191 & 72742114 \end{pmatrix}$$

由 $\hat{\sum} = \frac{1}{5+5-2}(S_1 + S_2)$ 可得

$$\hat{\sum}^{-1} = \begin{pmatrix} 7.36583 & -2.39931 & 1.80795 \\ -2.39931 & 0.982918 & -0.870921 \\ 1.80795 & -0.870921 & -0.932718 \end{pmatrix} \times 10^{-6}$$

(3) 求线性判别函数 $W(X)$

$$\alpha = \hat{\sum}^{-1}(\overline{X}^{(1)} - \overline{X}^{(2)}) = (2.384025 \quad -0.0490255 \quad -1.348214)' \times 10^{-3}$$

$$W(X) = a'(X - \overline{X}) = a'(X - \frac{1}{2}(\overline{X}^{(1)} + \overline{X}^{(2)}))$$

$$W(X) = (2.384025x_1 - 0.0490255x_2 - 1.348214x_3 + 9971.38) \times 10^{-3}$$

(4) 对已知类别的样品判别分类

样品省份	$W(X)$ 的值	判归类别
河 北	5.951866908	1
山西	6.436053832	1
辽宁	4.1363796	1
内蒙古	6.595714095	1
黑龙江	7.616140366	1
上海	−3.368041416	2
江苏	−8.13631284	2
浙江	−3.750584108	2
山东	−1.852917887	2
广东	−13.62829855	2

(5) 对判别效果做检验

判别分析是假设两组样品取自不同的总体，如果两个总体的均值向量在统计上差异不

显著，做判别分析意义就不大。所谓判别效果的检验就是检验两个正态总体均值向量是否相等，根据第一章的相关知识可知检验的统计量为：

$$F=\frac{(n_1+n_2-2)-p+1}{(n_1+n_2-2)p}T^2\sim F(p,n_1+n_2-p-1)$$

其中 $T^2=(n_1+n_2-2)\left[\sqrt{\frac{n_1n_2}{n_1+n_2}}(\overline{X}^{(1)}-\overline{X}^{(2)})'S^{-1}\sqrt{\frac{n_1n_2}{n_1+n_2}}(\overline{X}^{(1)}-\overline{X}^{(2)})\right]$

将计算结果代入统计量，求得的统计 $p=0.017709435$，故在 $\alpha=0.05$ 的检验水平下，两总体差异显著，即判别函数有效。

(3) 对待判别样品判别归类的结果如下：

样品省份	$W(X)$ 的值	判归类别
安徽	8.123868562	1
云南	8.556297141	1
福建	4.965703818	1

简单分析：回代率为百分之百，与资料符合，而待判的三个样本的判别结果表明：安徽、云南、福建为经济欠发达的省份，即第一类，结果符合实际。

5.3 贝叶斯(Bayes)判别法

从上节看距离判别法虽然简单，便于使用，但是该方法也有它明显的不足之处。第一，判别方法与总体各自出现的概率的大小无关；第二，判别方法与错判之后所造成的损失无关。Bayes 判别法就是为了解决这些问题而提出的一种判别方法。

5.3.1 基本思想

贝叶斯判别法对多个总体的判别不是考虑建立判别式，而是计算新给样品属于个总体的概率 $P(i/x)$，$i=1,\cdots,m$，比较这 m 个概率的大小，然后将新样品判归为来自概率最大的总体。

设有 m 个总体，$G_1,G_2,\cdots,G_m$，它们的先验概率分别为 $q_1,q_2,\cdots,q_m$，密度函数为 $f_1(x),f_2(x),\cdots,f_m(x)$，在观测到一个新样品 x 的情况下，可用贝叶斯公式计算它来自第 i 个总体的后验概率：

$$P(i/x)=\frac{q_if_i(x)}{\sum_{j=1}^{m}q_jf_j(x)},\quad i=1,2,\cdots,m$$

并且当

$$P(h/x)=\underset{1\leqslant i\leqslant m}{\mathrm{Max}}P(i/x)$$

时，判定 x 来自第 h 个总体。

另外，有时为了合理考虑错判所带来的损失，还使用错判损失最小的概念确定判别函数，这时，把 x 错判给第 h 个总体的平均损失定义为：

$$E(h/x)=\sum_{i\neq h}\frac{q_if_i(x)}{\sum_{j=1}^{m}q_jf_j(x)}L(h/i)$$

其中$L(h/i)$称为损失函数，它表示本来是第i个总体的样品错判为第h个总体的损失。于是建立判别准则为，如果

$$E(h/x)=\underset{1\leqslant i\leqslant m}{\mathrm{Min}}E(g/x)$$

则，判定x来自第h个总体。

显然考虑损失函数更为合理，但是由于实际应用中，由于$L(h/i)$不容易确定，经常在数学模型中假定各种错判的损失皆相等，这样，寻找h使后验概率最大实际上等价于使错判损失最小，则：

$$P(h/x)\xrightarrow{h}\mathrm{Max}\Leftrightarrow E(h/x)\xrightarrow{h}\mathrm{Min}$$

根据上述思想，在假定协方差矩阵相等的条件下，即可以导出判别函数。

5.3.2 多元正态总体的 Bayes 判别法

在实际问题中遇到的许多总体往往服从正态分布，下面给出p元正态总体的 Bayes 判别法，以及判别函数的导出。

(1) 待判样品的先验概率和密度函数

对于先验概率，一般可用样品频率来代替，即令$q_i=\dfrac{n_i}{n}$，其中n_i为用于建立判别函数的已知分类数据中来自第i总体样品的数目，且$n_1+n_2+\cdots+n_m=n$，或者干脆令先验概率相等，即$q_i=\dfrac{1}{m}$，这时可以认为先验概率不起作用。

对于第i总体的密度函数，设p元正态分布密度函数为：

$$f_i(x)=(2\pi)^{-\frac{p}{2}}\left|\sum^{(i)}\right|^{-\frac{1}{2}}\cdot\exp\left\{-\frac{1}{2}(x-\mu^{(i)})'\sum^{(i)-1}(x-\mu^{(i)})\right\}$$

式中$\mu^{(i)}$和$\sum^{(i)}$分别是第i总体的均值向量(p维)和协差阵(p阶)。把$f_i(x)$代入$P(i/x)$的表示式中，因为我们只关心寻找使$P(i/x)$的i，而分式中的分母不论i为何值都是常数，故可改令

$$q_if_i(x)\xrightarrow{i}\mathrm{Max}$$

对$q_if_i(x)$取对数并去掉与i无关的项，记为

$$\begin{aligned}Z(i/x)&=\ln q_i-\frac{1}{2}\ln\left|\sum^{(i)}\right|-\frac{1}{2}(x-\mu^{(i)})'\sum^{(i)-1}(x-\mu^{(i)})\\&=\ln q_i-\frac{1}{2}\ln\left|\sum^{(i)}\right|-\frac{1}{2}x'\sum^{(i)-1}x-\frac{1}{2}\mu^{(i)\prime}\sum^{(i)-1}\mu^{(i)}+x'\sum^{(i)-1}\mu^{(i)}\end{aligned}$$

则问题可化为　$Z(i/x)\xrightarrow{i}\max$

假定k个总体协方差阵相同，即$\sum^{(1)}=\sum^{(2)}=\cdots\sum^{(k)}=\sum$，这时$Z(i/x)$中$\dfrac{1}{2}\ln\left|\sum^{(i)}\right|$和$\dfrac{1}{2}x'\sum^{(i)-1}x$两项与$i$无关，求最大时可以去掉，最终得到如下形式的判别函数与判别准则

$$\begin{cases}y(i/x)=\ln q_i-\dfrac{1}{2}\mu^{(i)\prime}\sum^{-1}\mu^{(i)}+x'\sum^{-1}\mu^{(i)}\\ y(i/x)\xrightarrow{i}\mathrm{Max}\end{cases}$$

（2）计算后验概率

进行计算分类时，主要根据判别式 $y(i/x)$ 的大小，而它不是后验概率 $P(i/x)$，但是有了 $y(i/x)$ 之后，就可以根据下式算出后验概率 $P(i/x)$：

$$P(i/x)=\frac{\exp\{y(i/x)\}}{\sum_{i=1}^{m}\exp\{y(i/x)\}}$$

容易看出，$q_h f_h(x)=\max\limits_{1\leqslant i\leqslant m} q_i f_i(x)\Rightarrow P(h/x)=\max\limits_{1\leqslant i\leqslant k} P(i/x)$，则把样品 x 归为第 h 总体。

例 2 利用距离判别法中的例 1 的各省份的生产总值的数据做 Bayes 判别分析这里组数 $k=2$，变量个数 $p=3$，两类总体各有 5 个样品，$n_1=5$，$n_2=5$，有 3 个待判样品，为方便起见，假定两个总体协差阵相等。

两组线性判别的计算过程如下：

（1）
$$\overline{X}^{(1)}=\begin{pmatrix}1429.31\\7498.06\\5091.226\end{pmatrix}\qquad \overline{X}^{(2)}=\begin{pmatrix}1978.01\\17504.64\\14816.69\end{pmatrix},$$

$$q_1=q_2=\frac{5}{10}\qquad \ln q_1=\ln q_2=-0.693147$$

（2）计算样本协差阵，求出 $\hat{\sum}^{-1}$

$$\hat{\sum}^{-1}=\begin{pmatrix}7.36583 & -2.39931 & 1.80795\\ -2.39931 & 0.982918 & -0.870921\\ 1.80795 & -0.870921 & -0.932718\end{pmatrix}\times 10^{-6}$$

（3）求线性判别函数 $y(i/x)$

由 $y(i/x)=\ln q_i-\frac{1}{2}\mu^{(i)\prime}\sum^{-1}\mu^{(i)}+x'\sum^{-1}\mu^{(i)}\quad i=1,2$

$$f_1=0.001742585x_1-0.000493432x_2+0.000802588x_3-2.131683225$$
$$f_2=-0.00064144x_1-0.000444406x_2+0.002150802x_3-12.10305938$$

（4）将原各组样品进行回判的结果，并得待判别结果如下：

回判结果表明，总的回代判对率为 100%，这与统计资料的结果相符，并与前边的距离判别法的结果也相同。

G_1	f_1	f_2	回判类别
河北	2.768177011	−3.18369	1
山西	−1.00934013	−7.44539	1
辽宁	1.284957141	−2.85142	1
内蒙古	0.013025342	−6.58269	1
黑龙江	0.670124956	−6.94602	1
G_2			
上海	2.397743453	5.765785	2
江苏	5.310078378	13.44639	2
浙江	2.866435994	6.61702	2
山东	5.15309159	7.006009	2

续表

G1	f1	f2	回判类别
广东	7.120320867	20.74862	2
待判样品			
安徽	1.071047508	−7.05282	1
云南	0.53052524	−8.02577	1
福建	1.228263841	−3.73744	1

5.4 费舍(Fisher)判别法

费舍(*Fisher*)判别法的思想是投影，将 k 组 p 维数据投影到某一个方向，使得它们的投影组与组之间尽可能地分开，使投影每组内部离散性最小，该方法对总体的分布并未提出什么特定的要求。

具体的过程：针对 p 维空间中的某点 $x=(x_1,x_2,\cdots,x_p)$ 寻找一个能使它降为一维数值的线性函数 $y(x)$：

$$y=c_1x_1+c_2x_2+\cdots+c_px_p$$

其中 $c_1,c_2\cdots c_p$ 为待求的判别函数的系数，判别函数的系数的确定原则是使两投影组间区别最大，使每组内部离散性最小。有了判别函数后，对于一个新的样品，将 p 个指标的具体数值代入判别式中求出 y 值，然后与判别临界值进行比较，并判别其应属于哪一组。

5.4.1 两组判别分析

设有两组总体 G_1 和 G_2，相应抽出样品个数为 $n_1,n_2(n_1+n_2)=n$，每个样品观测 p 个指标得观测数据如下，

总体 G_1 的样本数据为：

$$\begin{matrix} X_1^{(1)} \\ X_2^{(1)} \\ \vdots \\ X_{n_1}^{(1)} \end{matrix}\begin{bmatrix} x_{11}^{(1)} & x_{12}^{(1)} & \cdots & x_{1p}^{(1)} \\ x_{21}^{(1)} & x_{22}^{(1)} & \cdots & x_{2p}^{(1)} \\ \cdots & \cdots & & \cdots \\ x_{n_11}^{(1)} & x_{n_12}^{(1)} & \cdots & x_{n_1p}^{(1)} \end{bmatrix}$$

该总体的样本指标平均值为：

$$\overline{X}^{(1)}=(\overline{x}_1^{(1)},\overline{x}_2^{(1)},\cdots,\overline{x}_p^{(1)})'$$

总体 G_2 的样本数据为：

$$\begin{matrix} X_1^{(2)} \\ X_2^{(2)} \\ \vdots \\ X_{n_2}^{(2)} \end{matrix}\begin{bmatrix} x_{11}^{(2)} & x_{12}^{(2)} & \cdots & x_{1p}^{(2)} \\ x_{21}^{(2)} & x_{22}^{(2)} & \cdots & x_{2p}^{(2)} \\ \cdots & \cdots & & \cdots \\ x_{n_21}^{(2)} & x_{n_22}^{(2)} & \cdots & x_{n_2p}^{(2)} \end{bmatrix}$$

该总体的样本指标平均值为：

$$\overline{X}^{(2)}=(\overline{x}_1^{(2)},\overline{x}_2^{(2)},\cdots,\overline{x}_p^{(2)})'$$

设判别函数

$$y(x)=c_1x_1+c_2x_2+\cdots+c_px_p$$

则 G_1 的数据对应的判别值为：

$$\begin{cases} y_1^{(1)}=c_1x_{11}^{(1)}+c_2x_{12}^{(1)}+\cdots+c_px_{1p}^{(1)} \\ y_2^{(1)}=c_1x_{21}^{(1)}+c_2x_{22}^{(1)}+\cdots+c_px_{2p}^{(1)} \\ \vdots \\ y_{n_1}^{(1)}=c_1x_{n_11}^{(1)}+c_2x_{n_12}^{(1)}+\cdots+c_px_{n_1p}^{(1)} \end{cases}$$

则 G_2 的数据对应的判别值为：

$$\begin{cases} y_1^{(2)}=c_1x_{11}^{(2)}+c_2x_{12}^{(2)}+\cdots+c_px_{1p}^{(2)} \\ y_2^{(2)}=c_1x_{21}^{(2)}+c_2x_{22}^{(2)}+\cdots+c_px_{2p}^{(2)} \\ \vdots \\ y_{n_2}^{(2)}=c_1x_{n_21}^{(2)}+c_2x_{n_22}^{(2)}+\cdots+c_px_{n_2p}^{(2)} \end{cases}$$

令

$$\bar{y}^{(1)}=\frac{1}{n_1}\sum_{i=1}^{n_1}y_i^{(1)},\bar{y}^{(2)}=\frac{1}{n_2}\sum_{i=1}^{n_2}y_i^{(2)}$$

根据费舍准则，要使判别的结果满足两组间区别最大，每组内部离散性最小。则判别函数的系数 $c_1,c_2\cdots c_p$ 应该能够使：

$$I=\frac{(\bar{y}^{(1)}-\bar{y}^{(2)})^2}{\sum_{i=1}^{n_1}(y_i^{(1)}-\bar{y}^{(1)})^2+\sum_{i=1}^{n_2}(y_i^{(2)}-\bar{y}^{(2)})^2}$$

取得最大值。根据微积分的知识，$c_1,c_2,\cdots,c_p$ 为方程组：

$$\frac{\partial I}{\partial c_i}=0,i=1,2,\cdots,p$$

的解。可以证明，最优判别函数系数 $c_1,c_2,\cdots,c_p$ 为下述方程的解：

$$\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{bmatrix}=\begin{bmatrix} \bar{x}_1^{(1)}-\bar{x}_1^{(2)} \\ \bar{x}_2^{(1)}-\bar{x}_2^{(2)} \\ \vdots \\ \bar{x}_p^{(1)}-\bar{x}_p^{(2)} \end{bmatrix}$$

其中：$s_{kl}=\sum_{i=1}^{n_1}(x_{ik}^{(1)}-\bar{x}_k^{(1)})(x_{il}^{(1)}-\bar{x}_l^{(1)})+\sum_{i=1}^{n_2}(x_{ik}^{(2)}-\bar{x}_k^{(2)})(x_{il}^{(2)}-\bar{x}_l^{(2)})$

所以

$$\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{bmatrix}=\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}^{-1}\begin{bmatrix} \bar{x}_1^{(1)}-\bar{x}_1^{(2)} \\ \bar{x}_2^{(1)}-\bar{x}_2^{(2)} \\ \vdots \\ \bar{x}_p^{(1)}-\bar{x}_p^{(2)} \end{bmatrix}$$

确定了判别函数以后，还需要确定判别临界值（分界点）y_0，在两总体先验概率相等的假设下，一般常取：

$$y_0=\frac{n_1\bar{y}^{(1)}+n_2\bar{y}^{(2)}}{n_1+n_2}$$

若有一判别的对象其数据为（$x_{01},x_{02},\cdots,x_{0p}$），则其判别值为

$y = c_1x_{01} + c_2x_{02} + \cdots + c_px_{0p}$

1) 当$\bar{y}^{(1)} > \bar{y}^{(2)}$时，若$y > y_0$，则判别该对象属于$G_1$，若$y < y_0$，判别该对象属于$G_2$

1) 当$\bar{y}^{(2)} > \bar{y}^{(1)}$时，若$y > y_0$，则判别该对象属于$G_2$，若$y < y_0$，判别该对象属于$G_1$

例 3 利用距离判别法中的例 1 的各省份的生产总值的数据做 *Fisher* 判别分析

(1) 建立判别函数

利用前边的计算结果，可得 *Fisher* 判别函数的系数

$$\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = S^{-1}\begin{pmatrix} \bar{x}_1^{(1)} - \bar{x}_1^{(2)} \\ \bar{x}_2^{(1)} - \bar{x}_2^{(2)} \\ \bar{x}_3^{(1)} - \bar{x}_3^{(2)} \end{pmatrix} = \begin{pmatrix} 298.003 \\ -6.12818 \\ -168.527 \end{pmatrix} \times 10^{-6}$$

所以判别函数为：

$$y = (298.003x_1 - 6.12818x_2 - 168.527x_3) \times 10^{-6}$$

(2) 计算判别临界值 y_0

$$\bar{y}^{(1)} = \sum_{k=1}^{3} c_k \bar{x}_k^{(1)} = -0.47802, \quad \bar{y}^{(2)} = \sum_{k=1}^{3} c_k \bar{x}_k^{(2)} = -2.01483$$

所以
$$y_0 = \frac{n_1 \bar{y}^{(1)} + n_2 \bar{y}^{(2)}}{n_1 + n_2} = -1.2464$$

(3) 判别准则

因为$\bar{y}^{(1)} > \bar{y}^{(2)}$，所以判别准则为

$$\begin{cases} 若\ y > y_0 & 判\ x \in G_1 \\ 若\ y < y_0 & 判\ x \in G_2 \\ 若\ y = y_0 & 待判 \end{cases}$$

(4) 将原各组样品进行回判的结果，并得待判别结果如下：

G_1	判别函数 y 的值	判归类号
河北	−0.502438656	1
山西	−0.44191529	1
辽宁	−0.729374569	1
内蒙古	−0.421957757	1
黑龙江	−0.294404473	1
G_2		
上海	−1.667427196	2
江苏	−2.263461124	2
浙江	−1.715245033	2
山东	−1.478036755	2
广东	−2.949959338	2
待判样品		
安徽	−0.230938449	1
云南	−0.176884877	1
福建	−0.625709042	1

回判结果表明，总的回代判对率为 100%，这与统计资料的结果相符，并与前边的距离判别法、*Bayes* 判别法的结果也相同。

5.4.2 多组别费舍判别法

设有k组总体$G_1,\cdots,G_k$，相应抽出样品个数为$n_1,\cdots,n_k(n_1+\cdots+n_k=n)$，每个样品观测 p 个指标得观测数据如下，

总体 G_i 的样本数据为：

$$\begin{matrix} X_1^{(i)} \\ X_2^{(i)} \\ \vdots \\ X_{n_i}^{(i)} \end{matrix}\begin{bmatrix} x_{11}^{(i)} & x_{12}^{(i)} & \cdots & x_{1p}^{(i)} \\ x_{21}^{(i)} & x_{22}^{(i)} & \cdots & x_{2p}^{(i)} \\ \cdots & \cdots & & \cdots \\ x_{n_i1}^{(i)} & x_{n_i2}^{(i)} & \cdots & x_{n_ip}^{(i)} \end{bmatrix} \quad i=1,2,\cdots,k$$

该总体 G_i 的样本指标平均值为：

$\bar{x}^{(i)}=(\bar{x}_1^{(i)},\bar{x}_2^{(i)}\cdots\bar{x}_p^{(i)})'$，样本协差阵：$s^{(i)}$

设判别函数

$$y(x)=c_1x_1+c_2x_2+\cdots+c_px_p\triangleq c'x$$

则 G_i 的数据对应的判别值为：

$$\begin{cases} y_1^{(i)}=c_1x_{11}^{(i)}+c_2x_{12}^{(i)}+\cdots+c_px_{1p}^{(i)} \\ y_2^{(i)}=c_1x_{21}^{(i)}+c_2x_{22}^{(i)}+\cdots+c_px_{2p}^{(i)} \\ \qquad\vdots \\ y_{n_i}^{(i)}=c_1x_{n_i1}^{(i)}+c_2x_{n_i2}^{(i)}+\cdots+c_px_{n_ip}^{(i)} \end{cases}$$

令 $\bar{y}^{(i)}=\dfrac{1}{n_i}\sum\limits_{t=1}^{n_i}y_t^{(i)}$，则根据求随机变量线性组合的均值和方差的性质可知，$y(x)$ 在 G_i 上的样本均值和样本方差为

$$\bar{y}^{(i)}=c'\bar{x}^{(i)},\sigma_i^2=c's^{(i)}c$$

记$\bar{x}$ 为总的均值向量，则 $\bar{y}=c'\bar{x}$。

在多总体情况下，Fisher 准则就是要选取系数向量 c，使

$$\lambda=\frac{\sum\limits_{i=1}^{k}n_i(\bar{y}^{(i)}-\bar{y})^2}{\sum\limits_{i=1}^{k}q_i\sigma_i^2}$$

达到最大，其中是 q_i 人为的正的加权系数，它可以取为先验概率，在这里取 $q_i=n_i-1$，根据相关的知识：

$$\frac{\partial\lambda}{\partial C}=0\Rightarrow Ac=\lambda Ec$$

其中 E 为组内离差阵，A 为总体之间样本的协差阵，即 $E=\sum\limits_{i=1}^{k}q_is^{(i)}$，$A=\sum\limits_{i=1}^{k}n_i(\bar{x}^{(i)}-\bar{x})(\bar{x}^{(i)}-\bar{x})'$。

这说明了λ及c恰好是矩阵A关于矩阵E的广义特征根及其对应的特征向量。由于一般都要求加权协差阵 E 是正定的，由代数知识可知，存在上式非零特征根：

$$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m>0\text{ 且 }m\leqslant\min(k-1,p)$$

1）只取一个判别函数，那么判别系数就是就是最大的非负特征根 λ_1 多对应的特征向量 c。若有一判别的对象其数据为$(x_{01},x_{02},\cdots,x_{0p})$，则其判别值为

$$y = c_1 x_{01} + c_2 x_{02} + \cdots + c_p x_{0p}$$

一种简单的分类方法：若 $|y(x) - \bar{y}^{(i)}| = \min\limits_{1 \leqslant j \leqslant k} |y(x) - \bar{y}^{(j)}|$ 则判 $x \in G_i$。

2）若取 $m_0(m_0 \leqslant m)$ 个判别函数，前 m_0 非负特征根 λ_l 及其对应的特征向量 $c^{(l)}$，

$$y_l(x) = c^{(l)'} x \qquad l = 1, \cdots, m_0$$

若 $D_\gamma^2 = \min\limits_{1 \leqslant j \leqslant k} D_i^2$，则判 $x \in G_\gamma$，其中 $D_i^2 = \sum\limits_{l=1}^{m_0} [y_l(x) - \bar{y}_l^{(i)}]^2 \lambda_l \qquad i = 1, \cdots, k$

最后，为了选取有效的判别函数，对于每个判别函数必须给出一个用以衡量判别能力的指标 p_l，衡量判别函数判别能力的指标定义为：

$$p_l = \frac{\lambda_l}{\sum\limits_{i=1}^{m} \lambda_i} \qquad l = 1, \cdots, m$$

则 $m_0(m_0 \leqslant m)$ 个判别函数的判别能力为：$\sum\limits_{l=1}^{m_0} p_l = \dfrac{\sum\limits_{l=1}^{m_0} \lambda_l}{\sum\limits_{i=1}^{m} \lambda_i}$，如果 m_0 达到某个人定的值（比如 80%）则就认为 m_0 个判别函数就够了。

贝叶斯方法一般多用于多组判别分析，贝叶斯判别方法的数学模型所要求的条件严格，它要求各组变量必须服从多元正态分布，各组的协方差矩阵相等，各组的均值向量有显著差异。而费舍判别法主要要求各组均值向量有显著差异即可。

5.5 逐步判别法

距离判别法、Bayes 判别法以及 Fisher 判别法等都是利用给定的全部变量来建立判别法则，但这些变量在判别式中所起的作用，一般来说是不同的，也就是说各变量在判别式中判别能力不同，有些可能起重要作用，有些可能不是很重要。实证发现，如果将一些判别能力不重要的变量保留在判别式中，不仅会增加计算量，而且会产生干扰影响判别效果，反之，如果将重要变量忽略了，也会影响判别的效果。逐步判别法就是在判别过程中不断的提取重要变量和剔除不重要变量，最终得到最佳的判别法则的过程。

5.5.1 引入和剔除变量所用的检验统计量

根据逐步判别分析的基本思想，进行判别分析需要解决两个关键的问题，一个是引入或剔除判别变量的依据和检验问题；另外则是判别函数的及时导出的问题。其中的理论基础又在于如何对判别变量在区别各个总体中是否提供附加信息的检验。为此这里先给出如何对判别变量在区别各个总体中是否提供附加信息进行检验的基础理论。

设有 k 组总体 $G_1, \cdots, G_k$，相应抽出样品个数为 $n_1, \cdots, n_k (n_1 + \cdots + n_k = n)$，每个样品观测 p 个指标得观测数据如下，

总体 G_i 的样本数据为：

$$\begin{matrix} X_1^{(i)} \\ X_2^{(i)} \\ \vdots \\ X_{n_i}^{(i)} \end{matrix} \begin{bmatrix} x_{11}^{(i)} & x_{12}^{(i)} & \cdots & x_{1p}^{(i)} \\ x_{21}^{(i)} & x_{22}^{(i)} & \cdots & x_{2p}^{(i)} \\ \cdots & \cdots & & \cdots \\ x_{n_i1}^{(i)} & x_{n_i2}^{(i)} & \cdots & x_{n_ip}^{(i)} \end{bmatrix} \quad i = 1,2,\cdots,k$$

该总体 G_i 的样本指标平均值为：

$$\overline{X}^{(i)} = (\overline{x}_1^{(i)}, \overline{x}_2^{(i)}, \cdots, \overline{x}_p^{(i)})'。$$

今作统计假设

$$H_0: \mu^{(1)} = \mu^{(2)} = \cdots = \mu^{(k)}$$

如果接受这个假设，说明建立的判别函数效果肯定不好，当原假设被否定时，说明这 k 个总体可以区分，建立判别函数有意义。根据第一章建立的检验统计量为

$$\Lambda_p = \frac{|E|}{|A+E|} \sim \Lambda_p(n-k, k-1)$$

其中，$E = \sum_{a=1}^{k}\sum_{i=1}^{n_a}(X_i^{(a)} - \overline{X}^{(a)})'(X_i^{(a)} - \overline{X}^{(a)})$，$A = \sum_{a=1}^{k} n_a(\overline{X}^{(a)} - \overline{X})'(\overline{X}^{(a)} - \overline{X})$

Λ_p 值越小，表明相同总体间的差异越小，相对地，样本间总的差异越大，即各总体间有较大差异。

5.5.2 逐步判别的原则

由于逐步判别法涉及的数学计算非常的复杂，在此就不作详细的介绍，只介绍逐步判别的基本过程。

(1) 在 $x_1, x_2, \cdots, x_m$(即 m 个自变量) 中，先选出一个自变量，它使维尔克斯统计量 $\Lambda_i(i = 1,\cdots,m)$ 达到最小。为了方便起见，不失一般性，假定挑选的变量次序是按自然的次序，即第 r 步正好选中 x_r，第一步选中 x_1，则有 $\Lambda_1 = \min\{\Lambda_i\}$，并考察 Λ_1 是否落入接受域，如不显著，则表明一个变量也选不中，不能判别分析；如显著，则进入下一步。

(2) 在未选中的变量中，计算它们与已选中的变量 x_1 配合的 Λ 值。选择使 $\Lambda_{1i}(2 \leqslant i \leqslant m)$ 达到最小的作为第二个变量。仿此，如已选入了个 r 变量，不妨设是 $x_1, x_2, \cdots, x_r$，则在未选中的变量中逐次选一个与它们配合，计算 $\Lambda_{1,2\cdots,r,l}(r < l \leqslant m)$，选择使上式子达到极小的变量作为第 $r+1$ 个变量。并检验新选的第 $r+1$ 个变量能否提供附加信息，如不能则转入(4)，否则转入(3)

(3) 在已选入的 r 个变量中，要考虑较早选中的变量中其重要性有没有较大的变化，应及时把不能提供附加信息的变量剔除出去。剔除的原则等同于引进的原则，例如在已进入的 r 个变量中要考察 $x_l(1 \leqslant l \leqslant r)$ 是否需要剔除，就是计算 $\Lambda_{l,1,2\cdots,l-1,l+1,\cdots r}$ 选择达到极小的 l，看是否显著，如不显著将该变量剔除，仍回到(3)，继续考察余下的变量是否需要剔除，如显著则回到(2)。

(4) 这时既不能选进新变量，又不能剔除已选进的变量，将已选中的变量建立判别函数。

无论用哪一种判别方法去判别样品的归属问题，均不可能永远的作出正确的判断，判断函数效果的验证方法有：(1) 自身验证(2) 外部数据验证(3) 样本二分法(4) 交互验证，读者可自行选择合适的验证方式。

本章思考与练习

1. 判别分析与聚类分析有何区别。

2. 试析距离判别法、贝叶斯判别法、Fisher 判别法的异同。

3. 下表是一个班同学的各科成绩，试用 K 聚类法把学生分成两类，然后用距离判别法建立判别函数，并根据此判别函数对原样本进行回判。

姓名	数学	物理	语文	政治
lxy	99.00	98.00	78.00	80.00
lwr	88.00	89.00	89.00	90.00
lgm	79.00	80.00	95.00	97.00
lm	89.00	78.00	81.00	82.00
hah	75.00	78.00	95.00	96.00
yxy	60.00	65.00	85.00	88.00
clf	79.00	87.00	50.00	51.00
wzb	75.00	76.00	88.00	89.00
cld	60.00	56.00	89.00	90.00
fj	100.00	100.00	85.00	84.00

第六章　主成分分析

在实际问题中，我们经常会遇到研究多个变量的问题，而且在多数情况下，多个变量之间常常存在一定的相关性。由于变量个数较多，再加上变量之间的相关性，势必增加了分析问题的复杂性。如何把多个变量综合为少数几个代表性的变量，使得这几个代表性变量既能够代表原始变量的绝大多数信息，又互不相关，并且在新的综合变量基础上，可以进一步的统计分析，这时就需要进行主成分分析(Principal Component Analysis)。

6.1　主成分分析的基本原理

设对某一事物的研究涉及 p 个指标，总体 X 是一个 p 维随机向量 $X=(X_1,X_2,\cdots,X_p)'$，$E(X)=\mu$，$D(X)=\sum$，$\sum$ 是实对称矩阵。

对 X 进行线性变换(线性变换简单，实践效果好)，可以形成新的综合变量，用 F 表示，新的综合变量可以由 $X=(X_1,X_2,\cdots,X_p)'$ 的线性组合表示：

$$\begin{cases} F_1 = u_{11}X_1 + u_{21}X_2 + \cdots + u_{p1}X_p \\ F_2 = u_{12}X_1 + u_{22}X_2 + \cdots + u_{p2}X_p \\ \cdots\cdots \\ F_p = u_{1p}X_1 + u_{2p}X_2 + \cdots + u_{pp}X_p \end{cases}$$

由于不同的线性变换得到的综合变量的统计特性不尽相同，为了使得综合变量能最大程度地包含原变量所代表的信息，又能保证新指标之间保持信息不重叠，我们总是希望 $F_i=u_i'X$ 的方差尽可能大且各 F_i 之间相互独立。

又由于

$$D(F_i)=D(u_i'X)=u_i'\sum_{ui},$$

对于任给的常数 a，有

$$D(au_i'X)=au_i'\sum u_i a=a^2u_i'\sum_{ui}$$

因此，对 u_i 不加限制时，可使 $D(F_i)$ 任意增大，问题将变得没有意义。我们将线性变换约束在下面的原则之下：

(1) 每个主成分系数平方和为 1 即：$u_{1i}^2+u_{2i}^2+\cdots+u_{pi}^2=1 \quad (i=1,2,\cdots,p)$

(2) 主成分之前互不相关 即：$\mathrm{Cov}(F_i,F_j)=0 \quad (i\neq j)$

(3) 主成分方差依次递减，即 $D(F_1)\geqslant D(F_2)\geqslant\cdots\geqslant D(F_p)$

则新变量指标 $F_1,F_2,\cdots,F_p$ 分别称为原变量指标 $X_1,X_2,\cdots,X_p$ 的第 1，第 2，……，第 p 主成分，

我们按照 $F_1, F_2, \cdots, F_p$ 的信息贡献程度即方差大小来选择适当的主成分代表原变量，通常只挑选前几个方差最大的主成分，从而达到简化系统结构，抓住问题实质的目的。

总之，综合指标（主成分）有以下几个特点：

(1) 主成分个数远远少于原有变量的个数

原有变量综合成少数几个因子之后，因子将可以替代原有变量参与数据建模，这将大大减少分析过程中的计算工作量。

(2) 主成分能够反映原有变量的绝大部分信息

因子并不是原有变量的简单取舍，而是原有变量重组后的结果，因此不会造成原有变量信息的大量丢失，并能够代表原有变量的绝大部分信息。

(3) 主成分之间应该互不相关

通过主成分分析得出的新的综合指标（主成分）之间互不相关，因子参与数据建模能够有效地解决变量信息重叠、多重共线性等给分析应用带来的诸多问题。

6.2　主成分分析的推导

在本节开始之前，对推导过程中引用的两个线性代数的定理先作简单的介绍。

定理 6.1 若 A 是 $p \times p$ 阶实对称阵，则一定可以找到正交阵 U，使得有

$$U^{-1}AU = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & \cdots & 0 \\ & & \ddots & \\ 0 & & & \lambda_p \end{bmatrix}$$

其中 $\lambda_1, \lambda_2, \cdots, \lambda_p$ 是 A 的特征根。

定理 6.2　若上述矩阵 A 的特征根所对应的单位特征向量为 $u_1, u_2, \cdots, u_p$，

令

$$U \triangleq (u_1, \cdots, u_p) = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}$$

则实对称矩阵 A 的属于不同特征根所对应的特征向量是正交的，即

$$u_i \cdot u_j = 0 \Rightarrow UU' = U'U = I。$$

6.2.1　从协方差出发求解总体主成分

设 $X = (X_1, X_2, \cdots, X_p)'$ 为一个 p 维随机向量，$E(X) = \mu, D(X) = \sum$，$\sum$ 是实对称矩阵，考虑如下的线性变换：

$$F = a_1X_1 + a_2X_2 + \cdots + a_pX_p \triangleq a'X$$

其中，$a = (a_1, a_2, \cdots, a_p)', X = (X_1, X_2, \cdots, X_p)'$。

求主成分的过程，就是寻求 a，使得 $D(a'X)$ 尽可能大，即使

$$\begin{aligned} D(a'X) &= E(a'X - E(a'X))(a'X - E(a'X))' \\ &= a'E(X - E(X))(X - E(X))'a \end{aligned}$$

$$= a' \sum a$$

达到最大值，且 $a'a = 1$。

(一) 总体主成分的求法

设 $\sum$ 的特征值为 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p \geqslant 0$，正交化单位特征向量为 $u_1, u_2, \cdots, u_p$，则第 k 个主成分表示为

$$F_k = u'_k X = u_{1k} X_1 + u_{2k} X_2 + \cdots + u_{pk} X_p$$

及
$$\begin{cases} D(F_k) = \lambda_k \\ \text{Cov}(F_j, F_k) \overset{j \neq k}{=} 0 \end{cases}$$

(二) 推导过程

令
$$U \triangleq (u_1, \cdots, u_p) = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}$$

由前面线性代数的定理可知：$U'U = UU' = I$，且

$$\sum = U \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & \cdots & 0 \\ & & \ddots & \\ 0 & & & \lambda_p \end{bmatrix} U' = \sum_{i=1}^{p} \lambda_i u_i u'_i$$

可得

$$a' \sum a = \sum_{i=1}^{p} \lambda_i a' u_i u'_i a = \sum_{i=1}^{p} \lambda_i (a' u_i)(a' u_i)' = \sum_{i=1}^{p} \lambda_i (a' u_i)^2$$

结合 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p > 0$，可得

$$a' \sum a \leqslant \lambda_1 \sum_{i=1}^{p} (a' u_i)^2 = \lambda_1 (a'U)(a'U)' = \lambda_1 a'UU'a \leqslant \lambda_1$$

又
$$u'_1 \sum u_1 = u'_1 \left(\sum_{i=1}^{p} \lambda_i u_i u'_i \right) u_1 = \left(\sum_{i=1}^{p} \lambda_i u'_1 u_i u'_i u_1 \right) = \lambda_1,$$

因此，$a = u_1$，使得 $D(a'X) = a' \sum a$ 达到最大值 λ_1，即 $D(u'_1 X) = u'_1 \sum u_1 = \lambda_1$，

可见，第一个主成分 $F_1 = u_{11} X_1 + u_{21} X_2 + \cdots u_{p1} X_p$ 且 $D(F_1) = \lambda_1$

同理 $D(u'_i X) = u'_i \sum u_i = \lambda_i$，令 $F_i = u'_i X$，则可得

$$\text{Cov}(F_i, F_j) = \text{Cov}(u'_i X, u'_j X) = u'_i \Sigma u_j = u'_i \left(\sum_{a=1}^{p} \lambda_a u_a u'_a \right) u_j = 0 \quad i \neq j$$

由上述的推导表明，$X_1, X_2, \cdots, X_p$ 的主成分就是以 $\sum$ 的单位特征向量为系数的线性组合，它们互不相关，其方差为 $\sum$ 的特征根。

由于 $\sum$ 的特征根 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p \geqslant 0$，所以有：$D(F_1) \geqslant D(F_2) \geqslant \cdots \geqslant D(F_p) \geqslant 0$，了解这一点也就明白为什么主成分的名次是按特征根取值大小的顺序排列的。

在解决问题时，一般不是取 p 个主成分，而是根据累计贡献率的大小取 k 个主成分，这

样就达到了降维的目的。

定义 6.1 总方差中属于第 i 主成分 F_i 的比例 $\lambda_i / \sum_{i=1}^{p} \lambda_i$ 称为主成分 F_i 的贡献率。

第一主成分 F_1 的贡献率最大，表明它解释原始变量 $X = (X_1, X_2, \cdots, X_p)'$ 的能力最强，而 $F_2, F_3, \cdots, F_p$ 的解释能力依次递减。

主成分分析的目的就是为了减少变量的个数，因而一般是不会使用所有 p 个主成分的，忽略一些带有较小方差的主成分将不会给总方差带来大的影响。

前 m 个主成分的贡献率之和 $\sum_{i=1}^{m} \lambda_i / \sum_{i=1}^{p} \lambda_i$ 称为主成分 $F_1, F_2, \cdots, F_m$ 的累计贡献率，它表明 $F_1, F_2, \cdots, F_m$ 解释 $X_1, X_2, \cdots, X_p$ 的能力。

通常取(相对于 p)较小的 m，使得累计贡献达到一个较高的百分比(如 80% ～ 90%)。此时，$F_1, F_2, \cdots, F_m$ 可用来代替 $X_1, X_2, \cdots, X_p$ 从而达到降维的目的，而信息的损失却不多。

定义 6.2 第 k 个主成分 F_k 与原始变量 X_i 的相关系数 $\rho(F_k, X_i)$ 称做因子负荷量。

因子负荷量是主成分解释中非常重要的解释依据，因子负荷量的绝对值大小刻画了该主成分的主要意义及其成因，因子负荷量的计算公式为

$$\rho(F_k, X_i) = u_{ki} \sqrt{\lambda_k} / \sqrt{\sigma_{ii}} \quad k, i = 1, 2, \cdots, p$$

从上式可知，因子负荷量与向量系数 u_{ki} 成正比，与 X_i 的标准差成反比关系，因此，不能把因子负荷量与向量系数混为一谈。在解释主成分的成因或是第 i 个变量对第 k 个主成分的重要性时，应当把因子负荷量和 F_k 与 X_i 的变换系数结合起来。

例 1 设 $X = (X_1, X_2, X_3)'$ 的协差矩阵为

$$\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

其特征值为 $\lambda_1 = 5.8284, \lambda_2 = 2, \lambda_3 = 0.1716$，其相应的单位特征向量为：

$$u_1 = \begin{pmatrix} -0.3827 \\ -0.9239 \\ 0 \end{pmatrix}, u_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, u_3 = \begin{pmatrix} -0.9239 \\ -0.3827 \\ 0 \end{pmatrix},$$

可得各主成分：
$$\begin{cases} F_1 = -0.3827X_1 + 0.9239X_2 + 0X_3 \\ F_2 = 0X_1 + 0X_2 + 1X_3 \\ F_3 = -0.9239X_1 - 0.3827X_2 + 0X_3 \end{cases}$$

前两个主成分的贡献率：$\dfrac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} = 97.875\%$，可取 F_1, F_2 可用来代替 X_1, X_2, X_3 从而达到降维的目的，而信息的损失却不多。

6.2.2 从相关阵出发求解总体主成分

在实际应用时，往往指标的量纲不同，所以在计算之前先消除量纲的影响，而将原始变量标准化，标准化的数学变换：

$$Z_i = \frac{X_i - \mu_i}{\sqrt{\sigma_{ii}}} \qquad i = 1, 2, \cdots, p$$

其中，μ_i, σ_{ii} 分别表示变量 X_i 的期望与方差。

令

$$\sum^{\frac{1}{2}}=\begin{bmatrix}\sqrt{\sigma_{11}} & 0 & \cdots & 0\\ 0 & \sqrt{\sigma_{22}} & \cdots & 0\\ \vdots & & \ddots & \\ 0 & & & \sqrt{\sigma_{pp}}\end{bmatrix}$$

那么,原始变量 X 的标准化:

$$Z=(\sum^{\frac{1}{2}})^{-1}(X-\mu)$$

我们对 $X_1,X_2,\cdots,X_p$ 的主成分分析转变为对 $Z_1,Z_2,\cdots,Z_p$ 的主成分分析,考虑如下的线性变换:

$$F=a_1Z_1+a_2Z_2+\cdots+a_pZ_p\triangleq a'Z$$

则可得

$$\begin{aligned}D(F)=D(a'Z)&=a'Cov(Z,Z)a\\&=a'(\sum^{\frac{1}{2}})^{-1}\sum(\sum^{\frac{1}{2}})^{-1}a\\&=a'Ra\end{aligned}$$

其中:$R=(\sum^{\frac{1}{2}})^{-1}\sum(\sum^{\frac{1}{2}})^{-1}=\begin{bmatrix}1 & \rho_{12} & \cdots & \rho_{1p}\\ \rho_{12} & 1 & \cdots & \rho_{2p}\\ \vdots & & \vdots & \vdots\\ \rho_{1p} & \rho_{2p} & \cdots & 1\end{bmatrix}$为 $X_1,X_2,\cdots,X_p$ 的相关阵。

由上面的变换过程,从 Z 的协方差出发求主成分的过程,实际上等同于从 $X_1,X_2,\cdots,X_p$ 的相关阵出发求主成分,因此,由相关矩阵求主成分的过程与主成分个数的确定准则实际上是与由协差矩阵出发求主成分的过程与主成分个数的确定准则是相一致的,在此不再赘述。

若从 $X_1,X_2,\cdots,X_p$ 的相关阵出发求主成分,从 R 的表达形式,此时 $\sigma_{ii}=1$,因子负荷量

$$\rho(F_k,X_i)=u_{ki}\sqrt{\lambda_k}\quad k,i=1,2,\cdots,p$$

例 2 设 $X=(X_1,X_2,X_3)'$ 的协差矩阵为

$$\sum=\begin{bmatrix}16 & 2 & 30\\ 2 & 1 & 4\\ 30 & 4 & 100\end{bmatrix}$$

我们从相关阵出发求其主成分。

解:$X=(X_1,X_2,X_3)'$ 的相关阵为

$$R=\begin{bmatrix}1 & 0.5 & 0.75\\ 0.5 & 1 & 0.4\\ 0.75 & 0.4 & 1\end{bmatrix},$$

其特征值为 $\lambda_1=2.114,\lambda_2=0.646,\lambda_3=0.24$,

其相应的单位特征向量为:

$$u_3=\begin{pmatrix}0.7410\\ -0.1420\\ -0.6563\end{pmatrix},u_2=\begin{pmatrix}-0.2408\\ -0.8562\\ -0.4571\end{pmatrix},u_1=\begin{pmatrix}0.6269\\ 0.4967\\ 0.6002\end{pmatrix}$$

可得各主成分：

$$\begin{cases} F_1 = 0.7410X_1 + (-0.1420)X_2 + (-0.6563)X_3 \\ F_2 = -0.2408X_1 + (-0.8562)X_2 + 0.4571X_3 \\ F_3 = -0.6269X_1 + 0.4967X_2 + 0.6002X_3 \end{cases}$$

前两个主成分的贡献率：$\dfrac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} = 92\%$，可取 F_1，F_2 可用来代替 X_1，X_2，X_3 从而达到降维的目的，而信息的损失却不多。

实际分析过程中，我们可以从原始数据的协差阵出发，也可以从原始数据的相关矩阵从发，其求主成分的过程是一致的。但是，从协差阵出发和从相关阵出发所求得的主成分一般来说是有差别的，而且某种情况下差别还挺大。

一般而言，对于度量单位不同的指标或是取值范围彼此差异非常大的指标，我们不直接由其协差阵出发进行主成分分析，而应考虑将数据标准化，则从相关阵出发进行主成分分析。

6.2.3 样本的主成分

从前面求主成分的过程我们了解到，我们可以从协差阵或相关阵出发求得主成分。但在实际问题中，$\sum$ 或 R 一般都是未知的，此时，可用其估计值$\dfrac{S}{n-1}$（样本协差阵）来代替 $\sum$，用样本相关系数矩阵$\hat{R}$ 来代替 R。

总体 $X = (X_1, X_2, \cdots, X_p)$ 有 p 项指标（变量），抽取 n 个样品，每个样品测得 p 项指标（变量），资料矩阵为：

$$X \triangleq \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

则样本协差阵和样本相关阵为

$$\frac{S}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'$$

$$R = (r_{ij}), r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}}\ \sqrt{s_{jj}}}$$

样品的主成分分析，实际上就是从样本协差阵或样本相关阵出发求主成分的过程，由于两者数学过程一样，这里只介绍从样本相关阵出发求主成分的步骤。

主成分分析的主要步骤如下：

(1) 计算样本相关阵 R

$$\frac{S}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})'$$

$$R = (r_{ij}), r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}}\ \sqrt{s_{jj}}}$$

(2) 求出 R 的特征值 λ_i 及相应的正交化单位特征向量 u_i

R 的前 m 个较大的特征值 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m$，就是前 m 个主成分对应的方差，λ_i 对应的单位特征向量 u_i 就是主成分 F_i 的关于原变量的系数，则原变量的第 i 个主成分 F_i 为：$F_i =$

$u_i'X$,主成分的方差(信息)贡献率用来反映信息量的大小,α_i 为:$\alpha_i = \lambda_i / \sum_{i=1}^{m} \lambda_i$。

(3) 选择主成分

最终要选择几个主成分,即 $F_1, F_2, \cdots, F_m$ 中 m 的确定是通过方差(信息)累计贡献率 $G(m)$ 来确定

$$G(m) = \sum_{i=1}^{m} \lambda_i / \sum_{k=1}^{p} \lambda_k$$

当累积贡献率大于 85% 时,就认为能足够反映原来变量的信息了,对应的 m 就是抽取的前 m 个主成分。

(4) 计算主成分载荷

$$l_{ij} = \rho(F_i, X_j) = \sqrt{\lambda_i} u_{ij} (i, j = 1, 2, \cdots, p)$$

(这里从 X 的相关阵出发,$\sigma_{jj} = 1$)

例 3 下面我们根据表 1 给出的数据,对某农业生态经济系统做主成分分析。

x_1:人口密度(人/km^2)

x_2:人均耕地面积(hm^2)

x_3:森林覆盖(%)

x_4:农民人均纯收入(元/人)

x_5:人均粮食产量(kg/人)

x_6:经济作物占农作物播种面积比例(%)

x_7:耕地占土地面积比率(%)

x_8:果园与林地面积之比(%)

x_9:灌溉田占耕地面积之比(%)

表 1 某农业生态经济系统各区域单元的有关数据

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
363.912	0.352	16.101	192.11	295.34	26.724	18.492	2.231	26.262
141.503	1.684	24.301	1752.35	452.26	32.314	14.464	1.455	27.066
100.695	1.067	65.601	1181.25	270.12	18.266	0.162	7.474	12.489
143.739	1.336	33.205	1436.12	354.26	17.486	11.805	1.892	17.534
131.412	1.623	16.607	1405.69	586.59	40.683	14.401	0.303	22.932
68.337	2.032	76.204	1540.29	216.39	8.128	4.065	0.011	4.861
95.416	0.801	71.106	926.35	291.52	8.135	4.063	0.012	4.862
62.901	1.652	73.307	1501.24	225.25	18.352	2.645	0.034	3.201
86.624	0.841	68.904	897.36	196.37	16.861	5.176	0.055	6.167
91.394	0.812	66.502	911.24	226.51	18.279	5.643	0.076	4.477
76.912	0.858	50.302	103.52	217.09	19.793	4.881	0.001	6.165
51.274	1.041	64.609	968.33	181.38	4.005	4.066	0.015	5.402
68.831	0.836	62.804	957.14	194.04	9.11	4.484	0.002	5.79
77.301	0.623	60.102	824.37	188.09	19.409	5.721	5.055	8.413
76.948	1.022	68.001	1255.42	211.55	11.102	3.133	0.01	3.425
99.265	0.654	60.702	1251.03	220.91	4.383	4.615	0.011	5.593
118.505	0.661	63.304	1246.47	242.16	10.706	6.053	0.154	8.701

续表

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
141.473	0.737	54.206	814.21	193.46	11.419	6.442	0.012	12.945
137.761	0.598	55.901	1124.05	228.44	9.521	7.881	0.069	12.654
117.612	1.245	54.503	805.67	175.23	18.106	5.789	0.048	8.461
122.781	0.731	49.102	1313.11	236.29	26.724	7.162	0.092	10.078

解:(1) 求其相关系数矩阵,得表 2

表 2　相关系数矩阵

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
x_1	1	−0.327	−0.714	−0.336	0.309	0.408	0.79	0.156	0.744
x_2	−0.327	1	−0.035	0.644	0.42	0.255	0.009	−0.078	0.094
x_3	−0.714	−0.035	1	0.07	−0.74	−0.755	−0.93	−0.109	−0.924
x_4	−0.336	0.644	0.07	1	0.383	0.069	−0.046	−0.031	0.073
x_5	0.309	0.42	−0.74	0.383	1	0.734	0.672	0.098	0.747
x_6	0.408	0.255	−0.755	0.069	0.734	1	0.658	0.222	0.707
x_7	0.79	0.009	−0.93	−0.046	0.672	0.658	1	−0.03	0.89
x_8	0.156	−0.078	−0.109	−0.031	0.098	0.222	−0.03	1	0.29
x_9	0.744	0.094	−0.924	0.073	0.747	0.707	0.89	0.29	1

(2) 由相关系数矩阵计算特征值,以及各个主成分的贡献率与累计贡献率(表 3),由表 3 可知,第 1、第 2、第 3 主成分的累计贡献率已高达 86.596%(大于 85%),故只需要求出第 1、第 2、第 3 主成分 F_1,F_2,F_3 即可。

表 3　特征值及主成分贡献率

主成分	特征值	贡献率 /%	累计贡献率 /%
F_1	4.661	51.791	51.791
F_2	2.089	23.216	75.007
F_3	1.043	11.589	86.596
F_4	0.507	5.638	92.234
F_5	0.315	3.502	95.736
F_6	0.193	2.14	97.876
F_7	0.114	1.271	99.147
F_8	0.045 3	0.504	99.65
F_9	0.0315	0.35	100

(3) 对于特征值 $\lambda_1 = 4.661\,0$,$\lambda_2 = 2.089\,0$,$\lambda_3 = 1.0430$ 分别求因子载荷矩阵(SPSS 默认选项的输出)。

表 4　因子载荷矩阵

	F_1	F_2	F_3	占方差的百分数 /%
x_1	0.739	−0.532	−0.0061	82.918
x_2	0.123	0.887	−0.0028	80.191
x_3	−0.964	0.009 6	0.009 5	92.948

续表

	F_1	F_2	F_3	占方差的百分数/%
x_4	0.004 2	0.868	0.003 7	75.346
x_5	0.813	0.444	−0.0011	85.811
x_6	0.819	0.179	0.125	71.843
x_7	0.933	−0.133	−0.251	95.118
x_8	0.197	−0.1	0.97	98.971
x_9	0.964	−0.0025	0.009 2	92.939

(4) 在表 4 中每一列除于$\sqrt{\lambda_i}$，得到主成分分析的第 i 个主成分的系数。

表 5　主成分系数矩阵

	F_1	F_2	F_3	占方差的百分数/%
x_1	0.342298	−0.36808	−0.00597	82.918
x_2	0.056973	0.613698	−0.00274	80.191
x_3	−0.44652	0.006642	0.009302	92.948
x_4	0.001945	0.600552	0.003623	75.346
x_5	0.376575	0.307195	−0.00108	85.811
x_6	0.379354	0.123847	−0.122396	71.843
x_7	0.432158	−0.09202	−0.24577	95.118
x_8	0.091249	−0.06919	0.949794	98.971
x_9	0.446516	−0.00173	0.009008	92.939

分析：

(1) 第 1 主成分 F_1 与 x_1, x_5, x_6, x_7, x_9 呈现出较强的正相关，与 x_3 呈现出较强的负相关，而这几个变量则综合反映了生态经济结构状况，因此可以认为第 1 主成分 F_1 是生态经济结构的代表。

(2) 第 2 主成分 F_2 与 x_2, x_4, x_5 呈现出较强的正相关，与 x_1 呈现出较强的负相关，其中，除了 x_1 为人口总数外，x_2, x_4, x_5 都反映了人均占有资源量的情况，因此可以认为第 2 主成分 F_2 代表了人均资源量。

(3) 第 3 主成分 F_3 与 x_8 呈现出的正相关程度最高，其次是 x_3，而与 x_7 呈负相关，因此可以认为第 3 主成分在一定程度上代表了农业经济结构。

(4) 另外，表 4 中最后一列(占方差的百分数)，在一定程度上反映了 3 个主成分 F_1, F_2, F_3 包含原变量($x_1, x_2, \cdots, x_9$)的信息量多少。

显然，用 3 个主成分 F_1, F_2, F_3 代替原来 9 个变量($x_1, x_2, \cdots, x_9$)描述农业生态经济系统，可以使问题更进一步简化、明了。

最后，应当认识到主成分分析方法适用于变量之间存在较强相关性的数据，如果数据相关性较弱，运用主成分分析后不能起到很好的降维作用，即所得的各个主成分浓缩原始变量信息的能力差别不大。一般认为当原始数据大部分变量的相关系数都小于 0.3 时，运用主成分分析不会取得很好的效果。

本章思考与练习

1. 试述主成分分析的基本思想，主成分分析的作用体现在何处？
2. 试述根据协差阵进行主成分分析和根据相关阵进行主成分分析分区别。

第七章　因子分析

因子分析(Factor Analysis)是指研究从变量群中提取共性因子的统计技术。最早由英国心理学家C. E. 斯皮尔曼提出。他发现学生的各科成绩之间存在着一定的相关性,一科成绩好的学生,往往其他各科成绩也比较好,从而推想是否存在某些潜在的共性因子,或称某些一般智力条件影响着学生的学习成绩。因子分析可在许多变量中找出隐藏的具有代表性的因子。将相同本质的变量归入一个因子,可减少变量的数目,还可检验变量间关系的假设。

因子分析不仅仅可以用来研究变量之间的关系,还可以用来研究样品之间的相关关系,通常前者称之为 R 型因子分析,后者称之为 Q 型因子分析。本章着重介绍 R 型因子分析。

7.1　因子分析的基本理论

7.1.1　因子分析的数学模型

本节从一个例子开始,直观地引出因子分析的数学模型。

由50道题组成的一套综合素质测试卷,题目涉及:语言表达能力、逻辑思维能力、对事物的敏锐程度、思想修养、兴趣爱好、生活常识等方面。第 i 位应试者在各题上的得分 $(x_{i1},x_{i2},\cdots,x_{i50})$ 是可观测的,可看作一个50维变量 $(X_1,X_2,\cdots,X_{50})$ 的取值。每道题上的得分是表面现象,应试者在语言表达能力、逻辑思维能力、对事物的敏锐程度、思想修养、兴趣爱好、生活常识等方面(称公共因子)的能力大小才是本质的,但是这每个公共因子都比较抽象,是潜在的,难以直接加以观测或度量。我们希望充分利用应试者在各题上的得分 $(x_{i1},x_{i2},\cdots,x_{i50})$ 信息,分析计算出应聘者在每个公共因子方面的水平高低。这就是因子分析要解决的问题。

设有 m 个公共因子,由于它们是潜在且不可观测的,形式上记为 $(F_1,F_2,\cdots,F_m)$。假设第 $i(i=1,2,\cdots,50)$ 小题的测试分数 X_i 可表示为

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + \varepsilon_i, \quad (i=1,2,\cdots,p)$$

这意味着我们试图通过 m 个潜在的公共因子 $(F_1,F_2,\cdots,F_m)$ 来对第 i 小题的测试分数 X_i 线性地加以解释。

其中系数 $a_{i1},a_{i2},\cdots,a_{im}$ 称因子载荷,用来表达第 i 小题的测试分数 X_i 反映出的各公共因子方面的能力。

ε_i 表达了第 i 小题的测试分数 X_i 不能被 m 个公共因子线性解释的部分,称为特殊因子。特殊因子也不可观测,通常假定 $\varepsilon_i \sim N(0,\sigma_i^2)$,这里的 σ_i^2 作为特殊因子的方差,可理解为特殊因子的强度的度量。

为了便于研究，并消除由于观测值量纲的差异及数量级不同所造成的影响，将样本观测数据进行标准化处理，使标准化后的变量均值为0，方差为1。为方便起见把原始变量及标准化后的变量均用 X 表示。

根据前述的思路，给出 R 型因子分析的数学模型：

$$\begin{cases} X_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + \varepsilon_1 \\ X_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + \varepsilon_2 \\ \cdots\cdots \\ X_p = a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m + \varepsilon_p \end{cases}$$

引入矩阵记号：

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}, A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{pmatrix}, F = \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{pmatrix}$$

其中矩阵 A 称因子载荷矩阵。

模型可表达为

$$X = AF + \varepsilon$$

且满足：

(1)$m \leqslant p$；

(2)$\mathrm{Cov}(F,\varepsilon) = 0$　即 F 和 ε 是不相关的；

(3)$D(F) = \begin{pmatrix} 1 & & & 0 \\ & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{pmatrix} = I_m$ 即 $F_1, F_2, \cdots, F_m$ 不相关且方差皆为1，此时称因子模型为正交因子模型；

(4)$D(\varepsilon) = \begin{pmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_p^2 \end{pmatrix}$　即 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_m$ 不相关且方差不同。

因子分析的任务，首先是估计出 $\{a_{ij}\}$ 和方差 $\{\sigma_i^2\}$，然后将这些抽象的因子 $\{F_i\}$ 赋予有实际背景解释或说给以命名。最后，依据样品 p 项可观测指标值 $(x_{i1}, x_{i2}, \cdots, x_{ip})$，希望能测算出该样品在各公共因子上的水平高低(称因子得分)。

7.1.2　因子模型中的几个统计特征

为了便于对因子分析计算结果做解释，将因子分析数学模型中各个变量的统计意义加以说明是十分必要的。

假定因子模型中，各个变量以及公共因子、特殊因子都已经是标准化(均值为0，方差为1)的变量。

(1) 因子载荷的统计意义

已知模型：

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{ij}F_j + \cdots + a_{im}F_m + \varepsilon_i$$

则
$$\mathrm{Cov}(X_i, F_j) = \mathrm{Cov}(\sum_{j=1}^{m} a_{ij}F_j + \varepsilon_i, F_j) = a_{ij}$$

即 a_{ij} 是 X_i 与 F_j 的协方差，而注意到 X_i 与 F_j 都是均值为 0，方差为 1 的变量，因此，a_{ij} 同时也是 X_i 与 F_j 的相关系数。故因子载荷 a_{ij} 表示 X_i 依赖 F_j 的分量（比重），它反映了第 i 个变量在第 j 个公共因子上的相对重要性。

（2）变量共同度的统计意义

所谓变量 X_i 的共同度定义为因子载荷阵 A 中第 i 行元素的平方和，即

$$h_i^2 = \sum_{j=1}^{m} a_{ij}^2 \quad i = 1, \cdots, p$$

为了说明它的统计学意义，将下式两边求方差，即：

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{ij}F_j + \cdots a_{im}F_m + \varepsilon_i$$

$$\begin{aligned} D(X_i) &= a_{i1}^2 D(F_1) + a_{i2}^2 D(F_2) + \cdots + a_{ij}^2 D(F_j) + \cdots a_{im}^2 D(F_m) + D(\varepsilon_i) \\ &= h_i^2 + \sigma_i^2 \end{aligned}$$

由于 X 已经标准化，故有

$$1 = h_i^2 + \sigma_i^2$$

上式表明共同度 h_i^2 与剩余方差 σ_i^2 有互补的关系，h_i^2 越大表明 X_i 对公共因子的依赖程度越大，例如 $h_i^2 = 0.97$ 则说明 X_i 的 97% 的信息都被所选取的公共因子说明了。公共因子能解释 X_i 方差的比例越大，因子分析的效果也越好。

（3）公因子 F_j 的方差贡献的统计意义

将因子载荷阵 A 中各列元素的平方和记为

$$S_j = \sum_{i=1}^{p} a_{ij}^2 \quad j = 1, \cdots, p$$

称 S_j 为公共因子 F_j 对 X 的贡献，即 S_j 表示同一公共因子 F_j 对诸变量所提供的方差贡献之和，它是衡量公共因子相对重要性的指标。

7.2　因子载荷阵的估计方法

因子分析可以分为确定因子载荷，因子旋转及计算因子得分 3 个步骤，因子分析的首要任务是根据样本数据估计载荷矩阵 A。估计 A 的方法有很多种：主成分法、最小二乘法、极大似然法、α 因子提取法等。这里仅介绍使用较为普遍的主成分法。

设随机变量 $X = (X_1, \cdots, X_p)'$ 的协差阵为 $\sum$，$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p > 0$ 为 $\sum$ 的特征根，$u_1, u_2, \cdots, u_p$ 为对应的标准正交化特征向量，根据线性代数的知识有：

$$\sum = U \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix} U' = \sum_{i=1}^{p} \lambda_i u_i u_i'$$

$$= (\sqrt{\lambda_1} u_1, \quad \cdots, \quad \sqrt{\lambda_p} u_p) \begin{pmatrix} \sqrt{\lambda_1} u'_1 \\ \vdots \\ \sqrt{\lambda_p} u'_p \end{pmatrix}$$

若 $m \doteq p$，由因子模型：$X = AF + \varepsilon$，及 $D(F) = I_m$，可得

$$D(X) = D(AF) + D(\varepsilon) = AD(F)A' + D(\varepsilon) = AA' + D(\varepsilon)$$

综合上面两个协方差的式子，可得：

$$\begin{cases} \sum = (\sqrt{\lambda_1} u_1, \quad \cdots, \quad \sqrt{\lambda_p} u_p) \begin{pmatrix} \sqrt{\lambda_1} u'_1 \\ \vdots \\ \sqrt{\lambda_p} u'_p \end{pmatrix} \\ \sum = AA' + D(\varepsilon) \end{cases}$$

在变量共同度 $h_i^2 \quad i = 1, \cdots, p$ 都很大的情况下，可以假设 $D(\varepsilon) = 0$，那么我们可得 A 的一个估计：$A = (\sqrt{\lambda_1} u_1, \quad \cdots, \quad \sqrt{\lambda_p} u_p)$，也就是说除常数 $\sqrt{\lambda_j}$ 外，第 j 列因子的载荷恰是第 j 个主成分的系数 u_j，故称为主成分法。

在假设 $D(\varepsilon) = 0$ 的情况下我们已经得到因子载荷矩阵 A，然而，它实际上是毫无价值的，因为我们的目的是寻求用少数几个公共因子解释，故略去后面的 $p—m$ 项 $\lambda_{m+1} u_{m+1} u'_{m+1} + \cdots + \lambda_p u_p u'_p$ 对 $\sum$ 的贡献，于是得到：

$$\sum \approx (\sqrt{\lambda_1} u_1, \quad \cdots, \quad \sqrt{\lambda_m} u_m) \begin{pmatrix} \sqrt{\lambda_1} u'_1 \\ \vdots \\ \sqrt{\lambda_m} u'_m \end{pmatrix} = AA'$$

上式是假定了因子模型中特殊因子是不重要的，因而可以从 $\sum$ 的分解中忽略掉特殊因子的方差。

当 $\sum$ 未知，可用样本协差阵 $\frac{S}{n-1}$ 去代替，要经过标准化处理，则 $\frac{S}{n-1}$ 与相关阵 R 相同，仍然可做上面类似的表示。

一般设 $\hat{\lambda}_1 \geqslant \hat{\lambda}_2 \geqslant \cdots \geqslant \hat{\lambda}_p$ 为样本相关阵 R 的特征根，相应的标准正交化特征向量为 $\hat{u}_1, \hat{u}_2, \cdots, \hat{u}_p$，设 $m < p$，则因子载荷矩阵的估计 $\hat{A} = (\hat{a}_{ij})$ 即

$$\hat{A} = (\sqrt{\hat{\lambda}_1}\ \hat{e}_1, \quad \cdots, \quad \sqrt{\hat{\lambda}_m}\ \hat{e}_m)$$

那么如何确定公因子的数目 m 呢？一般而言，这取决于问题的研究者本人，对于同一问题进行因子分析时，不同的研究者可能会给出不同的公因子数。当然，有时候有数据的本身特征可以很明确地确定出因子的数目。当用主成分法进行因子分析时，也可以借鉴确定主成分个数的准则，如所选取的公因子的信息量的和达到总体信息量的一个合适的比例为止。但对这些准则不应生搬硬套，应具体问题具体分析，总之要使所选取的公因子能够合理地描述原始变量相关阵的结构，同时要有利因子模型的解释。

例 1 假定某地固定资产投资率 x_1，通货膨胀率 x_2，失业率 x_3，相关系数矩阵为：

$$\begin{bmatrix} 1 & 1/5 & -1/5 \\ 1/5 & 1 & 2/5 \\ -1/5 & 2/5 & 1 \end{bmatrix}$$

试用主成分分析法求因子分析模型。

解:特征根为:$\lambda_1 = 1.4, \lambda_2 = 1.1464, \lambda_3 = 0.4536$

$$U = \begin{bmatrix} 0 & 0.8881 & -0.4597 \\ 0.7071 & 0.3251 & 0.6280 \\ 0.7071 & -0.3251 & -0.6280 \end{bmatrix}$$

。

$$A = (\sqrt{\lambda_1}u_1, \quad \sqrt{\lambda_2}u_2, \quad \sqrt{\lambda_3}u_3)$$

$$= \begin{bmatrix} 0 & 0.9509 & -0.3096 \\ 0.8367 & 0.3481 & 0.4230 \\ 0.8367 & -0.3481 & -0.4230 \end{bmatrix}$$

$$x_1 = 0.9509F_2 - 0.3096F_3$$

$$x_2 = 0.8367F_1 + 0.3481F_2 + 0.4230F_3$$

$$x_3 = 0.8367F_1 - 0.3481F_2 - 0.4230F_3$$

$\dfrac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} = 84.88\%$,可取前两个因子 F_1 和 F_2 为公共因子,第一公因子 F_1 为物价就业因子,对 X 的贡献为 1.4。第二公因子 F_2 为投资因子,对 X 的贡献为 1.1464。

7.3 因子旋转

在主成分分析中,每个主成分相应的系数 a_{ij} 是唯一确定的,与此相反,在因子分析中,每个因子的相应系数不是唯一的,即因子载荷矩阵不是唯一的。

若 Γ 为任一 $m \times m$ 阶正交阵,则因子模型:$X = AF + \varepsilon$,可写成:

$$X = (A\Gamma)(\Gamma'F) + \varepsilon$$

且仍然满足约束条件:

(1)$D(\Gamma'F) = \Gamma'D(F)\Gamma = I_m$

(2)$\mathrm{Cov}(\Gamma'F, \varepsilon) = \Gamma'\mathrm{Cov}(F, \varepsilon) = 0$

所以,$\Gamma'F$ 也是公共因子,$A\Gamma$ 也是因子载荷阵。

因子载荷的不唯一性是一个非常有利的性质,因为我们建立因子分析的目的不仅仅是要找出公共因子以及对变量进行分组,更重要的要知道每个公共因子的意义,以便进行进一步的分析,如果每个公共因子的含义不清,则不便于进行实际背景的解释。由于因子载荷阵是不唯一的,所以可以对因子载荷阵进行旋转(用一个正交阵右乘 A)使因子载荷阵的结构简化,使载荷矩阵每列或行的元素平方值向 0 和 1 两极分化。也就是说使每个变量仅在一个公共因子上有较大载荷,而在其他公共因子上的载荷较小。这种变换因子载荷阵的方法称为因子轴的旋转,而旋转的方法有多种,如正交旋转、斜交旋转等,本节只介绍常用的方差最大正交旋转法。

首先考虑 $m = 2$ 的情形。

设因子载荷阵 $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ \vdots & \vdots \\ a_{p1} & a_{p2} \end{bmatrix}$

对 A 按行计算共同度 $h_i^2=\sum_{j=1}^{2}a_{ij}^2, i=1,\cdots,p$，考虑到各个变量 X_i 的共同度之间的差异所造成的不平衡，需对中的元素进行规格化处理，即

$$(a'_{ij})^2=a_{ij}^2/h_i^2$$

对于规格化后的矩阵，为书写方便仍记为 A，施行方差最大正交旋转。

设旋转矩阵为：

$$T=\begin{pmatrix}\cos\phi & -\sin\phi\\ \sin\phi & \cos\phi\end{pmatrix}$$

记
$$B=AT=A\begin{pmatrix}\cos\phi & -\sin\phi\\ \sin\phi & \cos\phi\end{pmatrix}$$

$$=\begin{bmatrix}a_{11}\cos\phi+a_{12}\sin\phi & -a_{11}\sin\phi+a_{12}\cos\phi\\ \vdots & \vdots\\ a_{p1}\cos\phi+a_{p2}\sin\phi & -a_{p1}\sin\phi+a_{p1}\cos\phi\end{bmatrix}$$

$$=\begin{bmatrix}b_{11} & b_{12}\\ \vdots & \vdots\\ b_{p1} & b_{p2}\end{bmatrix}$$

是 $Z=T'F$ 的因子载荷矩阵。

正如前面所讲，我们希望所得结果能使载荷矩阵的每一列元素的绝对值尽可能向 1 和 0 两极分化，即原始变量中一部分主要与第一因子有关，另一部分主要与第二因子有关，也就是要求 $(b_{11}^2,\ \cdots,\ b_{p1}^2)$，$(b_{12}^2,\ \cdots,\ b_{p2}^2)$ 这两组的方差尽量大。为此，正交旋转的角度必须满足使旋转后得到因子载荷阵的总方差 $V_1+V_2\triangleq V$ 达到最大值，即：

$$\begin{cases}V=V_1+V_2=\max\\ \dfrac{\partial V}{\partial\phi}=0\end{cases},$$

其中 $V_\alpha=\dfrac{1}{p}\sum_{i=1}^{p}(b_{i\alpha}^2)^2-\left(\dfrac{1}{p}\sum_{i=1}^{p}b_{i\alpha}^2\right)^2 \quad \alpha=1,2$

经过计算，可得旋转角度可按下面的公式求得：

$$\tan 4\phi=\frac{D-2ABp}{C-2/(A^2-B^2)/p}$$

其中：
$$\mu_j=\left(\frac{a_{j1}}{h_j}\right)^2-\left(\frac{a_{j2}}{h_j}\right)^2 \quad v_j=2\frac{a_{j1}a_{j2}}{h_j^2}$$

$$A=\sum_{j=1}^{p}\mu_j,\ B=\sum_{j=1}^{p}v_j$$

$$C=\sum_{j=1}^{p}(\mu_j^2-v_j^2),\ D=2\sum_{j=1}^{p}\mu_j v_j$$

如果公共因子多于两个，我们可以逐次对每两个进行上述的旋转，设公共因子数 $m>2$，

(1) 第一轮旋转，每次取两个，全部配对旋转，变换共需进行 $m(m-1)/2$ 次；

(2) 对第一轮旋转所得结果用上述方法继续进行旋转，得到第二轮旋转结果，每一次旋转后，矩阵各列平方的相对方差之和总会比上一次有所增加；

(3) 当总方差的改变不大时，就可以停止旋转。

7.4 因子得分

因子分析模型建立后，还有一个重要的作用是应用因子分析模型去评价每个样品在整个模型中的地位，即进行综合评价。例如在前面 7.1 节的例子中，通过学生的成绩，我们希望了解学生的语言表达能力、逻辑思维能力、对事物的敏锐程度等因子的量化值，这时需要将公共因子用变量的线性组合来表示。

因子分析的数学模型为：

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2m} \\ \vdots & \vdots & & \vdots \\ \alpha_{p1} & \alpha_{p2} & \cdots & \alpha_{pm} \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix}$$

原变量被表示为公共因子的线性组合，当载荷矩阵旋转之后，公共因子可以做出解释，通常的情况下，我们还想反过来把公共因子表示为原变量的线性组合。

因子得分函数：

$$F_j = \beta_{j1} X_1 + \cdots + \beta_{jp} X_p \quad j = 1,2,\cdots,m$$

可见，要求每个因子的得分，必须求得分函数的系数，而由于 $p > m$，所以不能得到精确的得分，只能通过估计。

估计因子得分有很多种方法，如加权最小二乘法，回归法等。下面仅介绍回归法，它是 1939 年由 Thomson 提出来的，所以又称为汤姆森回归法。

Thomson 假设公共因子可以对 p 个变量做回归，由于假设变量及公共因子都已经标准化了，所以常数项为 0。即回归方程为：

$$\hat{F}_j = b_{j1} X_1 + \cdots + b_{jp} X_p \qquad j = 1,\cdots,m$$

令
$$B = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \vdots \\ \boldsymbol{b}_m \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mp} \end{bmatrix}, \text{则} \quad \hat{F} = BX$$

我们现在仅知道由样本值可得因子载荷阵 A，由因子载荷的意义知：

$$\alpha_{ij} = \gamma_{x_i F_j} = E(X_i F_j) = E[X_i(b_{j1} X_1 + \cdots + b_{jp} X_p)] = b_{j1}\gamma_{i1} + \cdots + b_{jp}\gamma_{ip}$$

$$= [\gamma_{i1} \quad \gamma_{i2} \quad \cdots \quad \gamma_{ip}] \begin{bmatrix} b_{j1} \\ b_{j2} \\ \vdots \\ b_{jp} \end{bmatrix}$$

则，我们有如下的方程组：

$$\begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1p} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2p} \\ \vdots & \vdots & & \vdots \\ \gamma_{p1} & \gamma_{p2} & \cdots & \gamma_{pp} \end{bmatrix} \begin{bmatrix} b_{j1} \\ b_{j2} \\ \vdots \\ b_{jp} \end{bmatrix} = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{pj} \end{bmatrix}$$

则可得：$RB' = A \Rightarrow B = A'R^{-1}$

其中

$$R = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1p} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2p} \\ \vdots & \vdots & & \vdots \\ \gamma_{p1} & \gamma_{p2} & \cdots & \gamma_{pp} \end{bmatrix} \text{为原始变量的相关系数矩阵}$$

$$A = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1m} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2m} \\ \vdots & \vdots & & \vdots \\ \alpha_{p1} & \alpha_{p2} & \cdots & \alpha_{pm} \end{bmatrix} \text{为载荷矩阵}$$

则

$$F = BX = A'R^{-1}X$$

其中 $X = (X_1, X_2, \cdots, X_p)'$，这就是估计因子得分的计算公式。

7.5 因子分析的步骤与逻辑框图

7.5.1 因子分析的步骤

设原始数据资料如下：

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n_1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

(1) 计算指标(变量) 的相关系数阵 R。

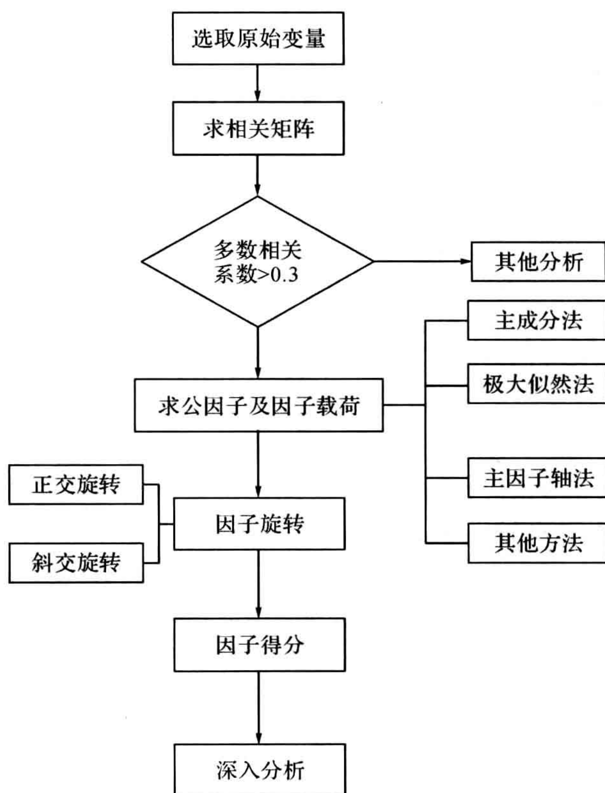
(2) 求 R 的特征根：$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p > 0$，$u_1, u_2, \cdots, u_p$ 为对应的单位正交化特征向量，可根据累计贡献率 $\sum_{i=1}^{m} \lambda_i / \sum_{i=1}^{p} \lambda_i \geqslant 85\%$，取前 m 个特征根，根据相应的单位特征向量得出因子载荷阵：

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{pmatrix} = \begin{pmatrix} u_{11}\sqrt{\lambda_1} & u_{12}\sqrt{\lambda_2} & \cdots & u_{1m}\sqrt{\lambda_m} \\ u_{21}\sqrt{\lambda_1} & u_{22}\sqrt{\lambda_2} & \cdots & u_{2m}\sqrt{\lambda_m} \\ \vdots & \vdots & & \vdots \\ u_{p1}\sqrt{\lambda_1} & u_{p2}\sqrt{\lambda_2} & \cdots & u_{pm}\sqrt{\lambda_m} \end{pmatrix}$$

(3) 根据 A 对 $\{F_i\}$ 的实际含义作出适当的解释，若含义不清，则对 A 做施行方差最大正交旋转。

(4) 计算因子得分。

7.5.2 因子分析的逻辑框图



本章思考与练习

1. 试述主成分分析与因子分析的联系与区别。
2. 因子分析主要应用于对哪些具体问题的分析?
3. 因子载荷 a_{ij} 的统计定义是什么?它在实际问题分析中的作用是什么?

第八章 典型相关分析

统计分析中，一般用简单相关系数反映两个变量之间的线性相关关系，用复相关系数反映一个变量与多个变量之间的线性相关关系。典型相关分(Canonical Correlation Analysis)是1936年由Hotelling在将线性相关性推广到两组变量的讨论中提出的，它是仿照主成分分析，把多变量与多变量之间的相关化为两个变量之间的相关分析，所揭示的是两组多元随机变量之间的关系。

设有两组变量，用 $X=(X_1,X_2,\cdots,X_p)'$，$Y=(Y_1,Y_2,\cdots,Y_q)'$ 表示，要研究两组变量的相关关系，一种方法是分别研究 X_i 和 Y_j 之间的相关关系，然后列出相关系数表进行分析，当两组变量较多时，这种做法不仅繁琐，也不易抓住问题的本质；另一种做法采用类似主成分分析的做法，在每一组变量中选择若干有代表性的综合指标(变量的线性组合)，通过研究两组的综合指标之间的关系来反映两组变量之间的相关关系。例如，在经济学中研究一组物品价格与消费量之间的关系，如猪肉和鸡蛋的价格分别用随机变量 X_1，X_2 来表示，猪肉与鸡蛋的消费量分别用随机变量 Y_1，Y_2 来表示，要研究随机变量 $X=(X_1,X_2)'$ 与 $Y=(Y_1,Y_2)'$ 的关系，从经济学观点就是希望构造一个 X_1，X_2 的线性函数 $U=a_{11}X_1+a_{12}X_2$ 称为价格指数及 Y_1，Y_2 的线性函数 $V=a_{21}Y_1+a_{22}Y_2$ 称为销售指数，要求它们之间具有最大相关性，这就是一个典型相关分析的问题。又如为了研究扩张性财政政策实施以后对宏观经济发展的影响，就需要考察有关财政政策的一系列指标，如财政支出总额的增长率、财政赤字增长率、国债发行额的增长率、税率降低率等与经济发展的一系列指标如国内生产总值增长率、就业增长率、物价上涨率等两组变量之间的相关程度。这时，常令随机向量 $X=(X_1,X_2,\cdots,X_p)'$ 表示 p 个财政政策指标，令 $Y=(Y_1,Y_2,\cdots,Y_q)'$ 为 q 个宏观经济的指标，构造综合的财产政策指标，综合的宏观经济指标，$U=a_{11}X_1+a_{12}X_2+\cdots+a_{1p}X_p$，$V=a_{21}Y_1+a_{22}Y_2+\cdots+a_{2q}Y_q$，研究扩张性财政政策实施以后对宏观经济发展的影响，就是研究随机 U 与 V 之间的关系。

典型相关分析的基本思想和主成分分析非常相似。首先在每组变量中找出变量的一个线性组合，使得两组的线性组合之间具有最大的相关系数。然后选取相关系数仅次于第一对线性组合并且与第一对线性组合不相关的第二对线性组合，如此继续下去，直到两组变量之间的相关性被提取完毕为止。因此，典型相关分析是把原来两组变量之间的相关，转化为研究从各组中提出的少数几个典型变量之间的典型相关，从而减少研究变量的个数。被选出的线性组合配对称为典型变量，它们的相关系数称为典型相关系数。典型相关系数度量了这两组变量之间联系的强度。

8.1 典型相关分析的数学描述

设有两组随机向量，X 代表第一组的 p 个变量，Y 代表第二组的 q 个变量，假设 $p \leqslant q$，令

$$D(X) = \sum_{11}, D(Y) = \sum_{22}, \mathrm{Cov}(X,Y) = \sum_{12} = \sum_{21}'$$

$$Z_{(p+q)\times 1} = \begin{bmatrix} X \\ \cdots \\ Y \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \\ \cdots \\ Y_1 \\ Y_2 \\ \vdots \\ Y_q \end{bmatrix}, \text{则 } D(Z) = \sum = \begin{pmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{pmatrix}$$

根据典型相关分析的基本思想，要进行两组随机向量间的相关分析，首先要计算出各组变量的线性组合 —— 典型变量，并使其相关系数达到最大。因此，我们设两组变量的线性组合分别为：

$$U = l_1 X_1 + l_2 X_2 + \cdots + l_p X_p \triangleq l'X$$

$$V = m_1 Y_1 + m_2 Y_2 + \cdots + m_q Y_q \triangleq m'Y$$

其中，$l = (l_1, l_2, \cdots, l_p)'$，$m = (m_1, m_2, \cdots, m_q)'$ 为任意非零常数向量，易见：

$$D(U) = D(l'X) = l'\mathrm{Cov}(X,X)l = l'\sum_{11} l$$

$$D(V) = D(m'Y) = m'\mathrm{Cov}(Y,Y)m = m'\sum_{22} m$$

$$\mathrm{Cov}(U,V) = l'\mathrm{Cov}(X,Y)m = l'\sum_{12} m$$

$$\rho_{UV} = \frac{l'\sum_{12} m}{\sqrt{l'\sum_{11} l}\sqrt{m'\sum_{22} m}}$$

我们希望寻找使相关系数达到最大的向量 l 与 m，由于随机向量乘以常数时并不改变它们的相关系数，所以，为防止结果的重复出现，令

$$D(U) = l'\sum_{11} l = 1$$

$$D(V) = m'\sum_{22} m = 1$$

那么，$\rho_{UV} = l'\sum_{12} m$，于是，我们的问题就成为在约束条件：$D(U) = l'\sum_{11} l = 1$，$D(V) = m'\sum_{22} m = 1$ 下，寻求 l, m 使 $\rho_{UV} = l'\sum_{12} m$ 达到最大。

8.2 总体典型相关

在约束条件：$D(U) = l'\sum_{11} l = 1$，$D(V) = m'\sum_{22} m = 1$ 下，寻求 l, m 使 $\rho_{UV} =$

$l'\sum_{12}m$ 达到最大，根据条件极值的求法引入 *Lagrange* 乘数，将问题转化为求：

$$\varphi(l,m)=l'\sum_{12}m-\frac{\lambda}{2}(l'\sum_{11}l-1)-\frac{\nu}{2}(m'\sum_{22}m-1) \tag{1}$$

的极大值，其中 λ,ν 是 Lagrange 乘数。

根据求极值的必要条件得

$$\begin{cases}\dfrac{\partial\varphi}{\partial l}=\sum_{12}m-\lambda\sum_{11}l=0\\ \dfrac{\partial\varphi}{\partial m}=\sum_{21}l-\nu\sum_{22}m=0\end{cases} \tag{2}$$

将上式分别左乘 l' 和 m'，则得：

$$\begin{cases}l'\sum_{12}m-\lambda l'\sum_{11}l=0\\ m'\sum_{21}l-\nu m'\sum_{22}m=0\end{cases} \tag{3}$$

由约束条件：$l'\sum_{11}l=1,m'\sum_{22}m=1$ 则有

$$\begin{cases}l'\sum_{12}m=\lambda\\ m'\sum_{21}l=\nu\end{cases} \tag{4}$$

而因为 $(l'\sum_{12}m)'=m'\sum_{21}l$，所以 $\lambda=\nu=l'\sum_{12}m$，且知 $\lambda=\rho_{UV}$，用 λ 代替方程组中的 ν，则方程组(2) 写为：

$$\begin{cases}\sum_{12}m-\lambda\sum_{11}l=0\\ \sum_{21}l-\lambda\sum_{22}m=0\end{cases} \tag{5}$$

由(5) 的第二个式子，我们可得到：

$$m=\frac{1}{\lambda}\sum\nolimits_{22}^{-1}\sum\nolimits_{21}l \tag{6}$$

将(6) 代入(5) 的第一个式子，得：

$$\sum\nolimits_{12}\frac{1}{\lambda}\sum\nolimits_{22}^{-1}\sum\nolimits_{21}l-\lambda\sum\nolimits_{11}l=0 \tag{7}$$

即有：

$$\sum\nolimits_{12}\sum\nolimits_{22}^{-1}\sum\nolimits_{21}l-\lambda^2\sum\nolimits_{11}l=0 \tag{8}$$

用 $\sum_{11}^{-1}$ 左乘(8) 式，得

$$\sum\nolimits_{11}^{-1}\sum\nolimits_{12}\sum\nolimits_{22}^{-1}\sum\nolimits_{21}l-\lambda^2 l=0 \tag{9}$$

同理可得

$$\sum\nolimits_{22}^{-1}\sum\nolimits_{21}\sum\nolimits_{11}^{-1}\sum\nolimits_{12}m-\lambda^2 m=0 \tag{10}$$

记

$$A=\sum\nolimits_{11}^{-1}\sum\nolimits_{12}\sum\nolimits_{22}^{-1}\sum\nolimits_{21},B=\sum\nolimits_{22}^{-1}\sum\nolimits_{21}\sum\nolimits_{11}^{-1}\sum\nolimits_{12}$$

(9)(10) 可表示为：

$$\begin{cases}Al=\lambda^2 l\\ Bm=\lambda^2 m\end{cases} \tag{11}$$

说明 λ^2 既是 A 又是 B 的特征根，l,m 就是其相应于 A 和 B 的特征向量。

可以证明，A 和 B 的特征根和特征向量有如下性质：

1. A 和 B 具有相同的非零特征根，且相等的非零特征根数目等于 p

2. A 和 B 的特征根均在 $0 \sim 1$ 之间。

设已求得的 A、B 的特征根依次为：

$$\lambda_1^2 \geqslant \lambda_2^2 \cdots \geqslant \lambda_p^2 > 0$$

对应于 A 的特征向量为 $l_1,l_2,\cdots,l_p$，对应于 B 的特征向量为 $m_1,m_2,\cdots,m_p$，这些特征根及相应的特征向量满足：

$$\begin{cases} Al_i = \lambda_i^2 l_i \\ Bm_i = \lambda_i^2 m_i \end{cases} \tag{12}$$

因为 λ_i^2 的特征向量不唯一，所以由前面的约定可知，规格化的特征向量满足：

$$D(U) = l'\sum_{11} i = 1,$$

$$D(V) = m'\sum_{22} m = 1$$

再者，为了讨论方便，我们约定 $\lambda = \rho_{UV}$ 总是大于零的，这很容易办到，后边的实例会做相关的讨论。以后我们讨论的特征向量，均是满足以上约定的 A,B 相应的特征向量。

因为 $\lambda = \rho_{UV}$，求 ρ_{UV} 最大值也就是求 λ 的最大值，因此，只要取最大特征值 λ_1^2 的平方根 λ_1，则 U_1 和 V_1 即具有最大的相关系数。令 l_1,m_1 为 λ_1^2 对应于 A,B 相应的特征向量(已规格化)，这时，$U_1 = l_1'X$ 与 $V_1 = m_1'Y$ 即分别为 X 与 Y 的规格化的线性组合，且具有最大的相关系数 λ_1。

综上所述，有如下定义：

定义 8.1 在一切使方差为 1 的线性组合 $l'X$ 与 $m'Y$ 中，其中两者相关关系数最大的 $U_1 = l_1'X$ 与 $V_1 = m_1'Y$ 称为第一对典型相关变量，它们的相关系数 λ_1 称为第一典型相关系数。

一般地，在定义了 $i-1$ 对典型相关变量后，在一切使方差为 1 且与前 $i-1$ 对典型相关变量都不相关的线性组合 $U_i = l_i'X$ 与 $V_i = m_i'Y$ 中，其两者相关系数最大者称为第 i 对典型相关变量，其相关系数称为第 i 对典型相关系数。

由上述推导，我们进一步有：求 X 与 Y 的第 i 个典型相关系数即求方程(11)的第 i 个最大根，而第 i 对典型变量即为 $U_i = l_i'X$ 与 $V_i = m_i'Y$，其中 l_i,m_i 为方程(11)当 $\lambda = \lambda_i$ 时所求得的解。

我们不加证明地给出典型变量以下的两个性质：

1. 由 $X_1,X_2,\cdots,X_p$ 所组成的典型变量 $U_1,U_2,\cdots,U_p$ 互不相关，同样的，由 $Y_1,Y_2,\cdots,Y_p$ 所组成的典型变量 $V_1,V_2,\cdots,V_p$ 也互不相关，且它们的方差等于 1。即有：

$$D(U_k) = 1, \quad D(V_k) = 1 \quad (k = 1,2,\cdots,p)$$

$$Cov(U_i,U_j) = 0, \quad Cov(V_i,V_j) = 0 \quad (i \neq j)$$

2. 同一对典型变量 U_i 及 V_i 的相关系数为 λ_i，不同对的典型变量 U_i 及 $V_j(i \neq j)$ 间互不相关，即有

$$Cov(U_i,V_j) = \begin{cases} \lambda_i \neq 0 & (i = j, i = 1,2,\cdots,p) \\ 0 & (i \neq j) \\ 0 & (j > p) \end{cases}$$

8.3 样本典型相关

在实际分析应用中，总体的协差阵通常是未知的，往往需要从研究的总体中随机抽取一个样本，根据样本估计出总体的协差阵，并在此基础上进行典型相关分析。

设 $Z=\begin{bmatrix}X\\Y\end{bmatrix}$服从正态分布 $N_{p+q}(\mu,\sum)$，从该总体中抽取样本容量为 n 的样本，得到下列数据矩阵：

$$X=\begin{bmatrix}X_{11} & X_{12} & \cdots & X_{1p}\\ X_{21} & X_{22} & \cdots & X_{2p}\\ \vdots & \vdots & & \vdots\\ X_{n1} & X_{n2} & \cdots & X_{np}\end{bmatrix}$$

$$Y=\begin{bmatrix}Y_{11} & Y_{12} & \cdots & Y_{1q}\\ Y_{21} & Y_{22} & \cdots & Y_{2q}\\ \vdots & \vdots & & \vdots\\ Y_{n1} & Y_{n2} & \cdots & Y_{nq}\end{bmatrix}$$

样本均值向量

$$\overline{X}=\begin{bmatrix}\overline{X}\\ \overline{Y}\end{bmatrix}\text{其中}\ \overline{X}=\frac{1}{n}\sum_{a=1}^{n}X_a\overline{Y}=\frac{1}{n}\sum_{a=1}^{n}Y_a$$

样本协差阵

$$\hat{\sum}=\begin{bmatrix}\hat{\sum}_{11} & \hat{\sum}_{12}\\ \hat{\sum}_{21} & \hat{\sum}_{22}\end{bmatrix}$$

其中，$\hat{\sum}_{kl}=\dfrac{1}{n-1}\sum_{j=1}^{n}(X_j-\overline{X})(Y_j-\overline{Y})'$，　$k,l=1,2,k\geqslant l,\hat{\sum}_{12}=\hat{\sum}'_{21}$，

由此可得矩阵 A 和 B 的样本估计：

$$A=\hat{\sum}_{11}^{-1}\hat{\sum}_{12}\hat{\sum}_{22}^{-1}\hat{\sum}_{21}$$

$$B=\hat{\sum}_{22}^{-1}\hat{\sum}_{21}\hat{\sum}_{11}^{-1}\hat{\sum}_{12}$$

如前所述，求解$\hat{A}$ 和$\hat{B}$ 的特征根及其相应的特征向量，即可得到所要求的典型相关变量及其典型相关系数。

这里需要注意，若样本数据矩阵已经标准化处理，此时样本的协差阵就等于

样本的相关系数矩阵

$$\hat{R}=\begin{bmatrix}\hat{R}_{11} & \hat{R}_{12}\\ \hat{R}_{21} & \hat{R}_{22}\end{bmatrix}$$

由此可得矩阵 A 和 B 的样本估计：

$$\hat{A}=\hat{R}_{11}^{-1}\hat{R}_{12}\hat{R}_{22}^{-1}\hat{R}_{21}$$

$$\hat{B}=\hat{R}_{22}^{-1}\hat{R}_{21}\hat{R}_{11}^{-1}\hat{R}_{12}$$

求解A和B的特征根及相应的特征向量，即可得到典型变量及典型相关系数。此时相当于从相关矩阵出发计算典型变量。

8.4 典型相关系数的显著性检验

在作两组变量的典型相关分析之前，首先应检验两组变量是否相关，如果不相关，即$\mathrm{Cov}(X,Y)=0$，则讨论两组变量的典型相关就毫无意义。

因此，在用样本数据进行典型相关分析时应就两组变量的协差阵是否为零进行检验。即检验假设

$$H_0: \Sigma_{12}=0, \qquad H_1: \Sigma_{12}\neq 0$$

设$Z=\begin{bmatrix}X\\Y\end{bmatrix}$服从正态分布$N_{p+q}(\mu,\Sigma)$，根据随机向量的检验理论可知，用于检验的似然比统计量为

$$\Lambda_0=\frac{|\hat{\Sigma}|}{|\hat{\Sigma}_{11}|\,|\hat{\Sigma}_{22}|}=\prod_{i=1}^{p}(1-\hat{\lambda}_i^2)$$

其中，$\hat{\lambda}_i^2$是A和B的特征根，按大小排列为：$\hat{\lambda}_1^2\geqslant\hat{\lambda}_2^2\geqslant\cdots\geqslant\hat{\lambda}_p^2>0$，当$n\gg 1$时，巴特莱特(Bartlett)证明，当$H_0$成立时，$Q_0=-m\ln\Lambda_0$近似服从$\chi^2(f)$分布，其中$m=(n-1)-\frac{1}{2}(p+q+1)$，自由度$f=pq$。在给定的显著性水平$\alpha$下，当由样本计算的$Q_0>\chi_\alpha^2$临界值时，拒绝原假设，认为第一对典型变量$\hat{U}_1,\hat{V}_1$存在相关性，其相关系数为$\hat{\lambda}_1$，即至少可以认为第一个典型相关系数$\hat{\lambda}_1$是显著的。将它出去之后，再检验其余的$p-1$个典型相关系数的显著性，这时计算：

$$\Lambda_1=\prod_{i=2}^{p}(1-\hat{\lambda}_i^2)$$

则统计量$Q_1=-[n-2-0.5(p+q+1)]\ln\Lambda_1$近似服从自由度$(p-1)(q-1)$的卡方分布，如果$Q_1>\chi_\alpha^2$，则认为$\hat{\lambda}_2$显著，即第二对典型变量$\hat{U}_2,\hat{V}_2$存在相关性，以下逐个进行检验，直到某个$\hat{\lambda}_k$不显著时截止。

8.5 典型相关系数的步骤及实例

典型相关分析计算步骤

(一) 根据分析目的建立原始矩阵

原始数据矩阵

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} & y_{11} & y_{12} & \cdots & y_{1q} \\ x_{21} & x_{2} & \cdots & x_{2p} & y_{21} & y_{22} & \cdots & y_{2q} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} & y_{n1} & y_{n2} & \cdots & y_{nq} \end{bmatrix}$$

（二）对原始数据进行标准化变化并计算相关系数矩阵

$$R = \begin{bmatrix} \hat{R}_{11} & \hat{R}_{12} \\ \hat{R}_{21} & \hat{R}_{22} \end{bmatrix}$$

其中 R_{11}，R_{22} 分别为第一组变量和第二组变量的相关系数阵，R_{12} 为第一组变量和第二组变量的相关系数，并且有 $R_{12} = R'_{21}$。

（三）求典型相关系数和典型变量

计算矩阵 $\hat{A} = \hat{R}_{11}^{-1}\hat{R}_{12}\hat{R}_{22}^{-1}\hat{R}_{21}$ 以及矩阵 $\hat{B} = \hat{R}_{22}^{-1}\hat{R}_{21}\hat{R}_{11}^{-1}\hat{R}_{12}$ 的特征值和特征向量，并规格化特征向量，则分别得典型相关系数和典型变量。

（四）检验各典型相关系数的显著性

例 1 康复俱乐部对 20 名中年人测量了三个生理指标：体重（$x1$），腰围（$x2$），脉搏（$x3$）；三个训练指标：引体向上次数（$y1$），起坐次数（$y2$），跳跃次数（$y3$）。分析生理指标与训练指标的相关性。

表 1

样本＼变量	x_1	x_2	x_3	y_1	y_2	y_3
1	191	36	50	5	162	60
2	189	37	52	2	110	60
3	193	38	58	12	101	101
4	162	35	62	12	105	37
5	189	35	46	13	155	58
6	182	36	56	4	101	42
7	211	38	56	8	101	38
8	167	34	60	6	125	40
9	176	31	74	15	200	40
10	154	33	56	17	251	250
11	169	34	50	17	120	38
12	166	33	52	13	210	115
13	154	34	64	14	215	105
14	247	46	50	1	50	50
15	193	36	46	6	70	31
16	202	37	62	12	210	120
17	176	37	54	4	60	25
18	157	32	52	11	230	80
19	156	33	54	15	225	73
20	138	33	68	2	110	43

解:由表 1 数据,我们可得到样本协差阵为:

	x_1	x_2	x_3	y_1	y_2	y_3
x_1	579.14	65.36	−61.86	−48.32	−723.63	−272.18
x_2	65.36	9.74	−7.74	−8.88	−122.87	−29.87
x_3	−61.86	−7.74	49.39	5.455	96.445	12.27
y_1	−48.32	−8.88	5.455	26.5475	218.6025	127.665
y_2	−723.63	−122.87	96.445	218.6025	3718.848	2039.635
y_3	−272.18	−29.87	12.27	127.665	2039.635	2497.91

则我们可得到:

$$\hat{\sum}_{11}=\begin{bmatrix}579.14 & 65.36 & -61.86\\ 65.36 & 9.74 & -7.74\\ -61.86 & -7.74 & 49.39\end{bmatrix},\hat{\sum}_{22}=\begin{bmatrix}26.55 & 218.60 & 127.67\\ 218.60 & 3718.85 & 2039.64\\ 127.67 & 2039.64 & 2497.91\end{bmatrix}$$

$$\hat{\sum}_{12}=\begin{bmatrix}-48.32 & -723.63 & -272.18\\ -8.88 & -122.87 & -29.87\\ 5.46 & 96.45 & 12.27\end{bmatrix},\hat{\sum}_{21}=\begin{bmatrix}-48.32 & -8.88 & 5.46\\ -723.63 & -122.87 & 96.45\\ -272.18 & -29.87 & 12.27\end{bmatrix}$$

计算 $\hat{\sum}_{11}$,$\hat{\sum}_{22}$ 的逆矩阵,可得

$$\hat{\sum}_{11}^{-1}=\begin{bmatrix}0.00723237 & -0.047214 & 0.00165941\\ -0.047214 & 0.42549329 & 0.00754531\\ 0.00165941 & 0.00754531 & 0.02350784\end{bmatrix}$$

$$\hat{\sum}_{22}^{-1}=\begin{bmatrix}0.0732399 & -0.0040789 & -0.00041\\ -0.0040789 & 0.00071416 & -0.00037\\ -0.0004126 & -0.0003747 & 0.000727\end{bmatrix}$$

计算得:

$$\hat{A}=\hat{\sum}_{11}^{-1}\hat{\sum}_{12}\hat{\sum}_{22}^{-1}\hat{\sum}_{21}=\begin{bmatrix}-0.2459454 & -0.0551887 & 0.04651367\\ 4.498811 & 0.90714323 & -0.7392212\\ -0.0575041 & -0.0138964 & 0.01728371\end{bmatrix}$$

$$\hat{B}=\hat{\sum}_{22}^{-1}\hat{\sum}_{21}\hat{\sum}_{11}^{-1}\hat{\sum}_{12}=\begin{bmatrix}0.16178831 & 2.03428439 & 0.223085\\ 0.04076171 & 0.54877371 & 0.091339\\ -0.0328274 & -0.4227509 & -0.03208\end{bmatrix}$$

求得特征值为:$\lambda_1^2=0.632994993$,$\lambda_2^2=0.040214862$,$\lambda_3^2=0.005267145$。典型相关系数分别为:$\lambda_1=0.796$,$\lambda_2=0.201$,$\lambda_3=0.073$。

利用 *Matlab* 求得:$\lambda_i^2(i=1,2,3)$ 对应于$\hat{A}$ 的单位特征向量为:

$$l_1=(-0.0635,0.9978,-0.0166)',$$

$$l_2=(-0.202,0.9757,-0.085)',$$

$$l_3=(-0.036,0.7347,0.6775)',$$

可以注意到单位特征向量的元素的平方和为 1,但是,我们要求的约束条件是:

$D(U)=l'\sum_{11}l=1$,$D(V)=m'\sum_{22}m=1$,而根据上边的单位特征向量:$l_1'\sum_{11}l_1=$

3.889569，并不满足约束条件，必须对特征向量进行规格化，如 l_1 的规划化为：$l_1/\sqrt{3.889569}=(0.032198,-0.50593,0.008417)'$，仍记规格化后$\hat{A}$ 的特征向量为 l_i，则

$$l_1=(-0.032198,0.50593,-0.008417)'$$
$$l_2=(-0.07829,0.378175,-0.03287)'$$
$$l_3=(-0.00794,0.162137,0.149514)'$$

此时，规格化后的特征向量必然满足 $D(U)=l'\sum_{11}l=1,D(V)=m'\sum_{22}m=1$。

同样的，我们可得 $\lambda_i^2(i=1,2,3)$ 对应于$\hat{B}$ 的规格化的特征向量为：

$$m_1=(0.066,0.017,-0.014)',$$
$$m_2=(-0.071,0.002,0.021)'$$
$$m_3=(-0.245,0.020,-0.008)'$$

到目前为止，我们得到三对典型相关变量：$U_i=l_i'X$ 与 $V_i=m_i'Y,i=1,2,3$，我们不必全部接受者三对典型的相关变量，下边我们进行检验。

$$\begin{aligned}\Lambda_0&=\prod_{i=1}^{3}(1-\hat{\lambda}_i^2)\\&=(1-0.632994993)(1-0.040214862)(1-0.005267145)\\&=0.350390621\end{aligned}$$

$$\begin{aligned}Q_0&=-m\ln\Lambda_0=-\left[(n-1)-\frac{1}{2}(p+q+1)\right]\ln\Lambda_0\\&=-\left[(20-1)-\frac{1}{2}(3+3+1)\right]\ln\Lambda_0\\&=-15.5\ln\Lambda_0=16.255\end{aligned}$$

$Q_0<\chi_{0.05}^2(9)=16.91896016$，故在 $\alpha=0.05$ 下，生理指标与训练指标之间不存在相关性；而在 $\alpha=0.10$ 下，$Q_0>\chi_{0.10}^2(9)=14.68366$，生理指标与训练指标之间存在相关性，且第一对典型变量相关性显著。

继续检验：

$$\Lambda_1=\prod_{i=2}^{3}(1-\hat{\lambda}_i^2)=(1-0.040214862)(1-0.005267145)=0.954729811$$

$$\begin{aligned}Q_1&=-m\ln\Lambda_1=-\left[(n-1)-\frac{1}{2}(p+q+1)\right]\ln\Lambda_1\\&=-\left[(20-1)-\frac{1}{2}(3+3+1)\right]\ln\Lambda_0\\&=-15.5\ln\Lambda_1=0.718\end{aligned}$$

$Q_1<\chi_{0.10}^2(4)=7.779434$，故在 $\alpha=0.10$ 下，第二对典型变量间相关性不显著。

说明生理指标和训练指标之间只有一对典型变量，即：

$$U_1=-0.0322X_1+0.5059X_2-0.0084X_3$$
$$V_1=0.066Y_1+0.017Y_2-0.014Y_3$$

$\rho_{U_1V_1}=l_1'\sum_{12}m_1=-0.796=-\lambda_1$，为了叙述的方便，一般的软件都是给出两组变量正的相关系数，这里可做相应的调整，可对 l_1 或 m_1 其中一个向量乘以 -1，如可令：$l_1^*=-l_1=(0.032198,-0.50593,0.008417)'$，那么相关系数是正数的规格化的第一对典型变量为：

$$U_1 = 0.0322X_1 - 0.5059X_2 + 0.0084X_3$$
$$V_1 = 0.066Y_1 + 0.017Y_2 - 0.014Y_3$$

本章思考与练习

1. 简述典型相关分析的基本思想。

2. 请查找我国2010年各省(市、自治区)如下两组变量的数据,对两组数据进行典型相关分析,并对分析结果进行评述。

第一组变量:常住人口、人均GDP、固定资产投资、引进外国直接投资、R&D经费投入、教育经费支出;

第二组变量:GDP增长率、非农产业增加值占GDP比重、人均最终消费支出、出口总额。

第九章　对应分析

对应分析(correspondence analysis)又称为相应分析,也称 R-Q 型因子分析,是在因子分析的基础上发展起来的一种多元统计分析方法。它可以应用于定量数据的分析,也可以应用于定性数据的分析,通过分析定性变量构成的列联表来揭示变量之间的关系。在因子分析中人们通常只是分析原始变量的因子结构,找出决定原始变量的公共因子,从而使问题的分析简化和清晰。这种研究对象是变量的因子分析称为 R 型因子分析。但是对于有些问题来说,我们还需要研究样品的结构,若对于样品进行因子分析,称为 Q 型因子分析。当我们对同一观测数据同时施加 R 和 Q 型因子分析,并分别保留两个公共因子,则是对应分析的初步。

9.1　对应分析及基本思想

9.1.1　对应分析的数据类型

对应分析是一种描述性、探索性的数据分析方法,可以对定性与定量的数据类型进行分析:

(1) 定量数据:设有 n 个样品,每个样品有 p 项指标,原始资料阵为:

$$X \triangleq \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

(2) 定性数据:二维列联表的数据结构:

$n \times p$ 列联表

行＼列	A_1	A_2	…	A_p	合计
B_1	n_{11}	n_{12}	…	n_{1p}	n_1
B_2	n_{21}	n_{22}	…	n_{2p}	n_2
⋮	⋮	⋮	⋮	⋮	⋮
B_n .	n_{n1}	n_{n2}	…	n_{pp}	n_n
合计	n_1	n_2	…	n_p	n

虽然(1)(2) 的数据类型看起来不太相同,实际上它们做对应分析的出发样本矩阵是“相同”的,下边我们通过两个实例来讨论这个问题。

例 1 为了研究眼睛颜色与头发颜色之间关系，表 1 包含了 5387 名苏格兰北部的凯斯纳斯郡小学生的眼睛颜色与头发颜色，利用数据探讨眼睛颜色与头发颜色之间的对应关系。

表 1

眼睛颜色	头发颜色					合计
	金色	红色	棕色	深色	黑色	
深色	98	48	403	681	85	1315
棕色	343	84	909	412	26	1774
蓝色	326	38	241	110	3	718
浅色	688	116	584	188	4	1580
合计	1455	286	2137	1391	118	5387

例 2 用对应分析研究我国 2011 年部分省份的城镇居民家庭平均每人全年现金消费支出结构。

选取 7 个变量：X_1：食品支出 X_2：衣着支出 X_3：居住支出 X_4：家庭设备及服务支出 X_5：医疗保健支出 X_6：交通和通讯支出 X_7：文教娱乐、用品及服务支出。

样品为 10 个，即山西、内蒙古、辽宁、吉林、黑龙江、海南、四川、贵州、甘肃、青海。原始数据如下：

表 2 （单位：元）

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
山西	3558.04	1461.90	1327.78	832.74	1487.66	1419.43	851.30
内蒙古	4962.40	2514.09	1418.60	1162.87	2003.54	1812.07	1239.36
辽宁	5254.96	1854.63	1385.62	929.37	1899.06	1614.52	1208.30
吉林	4252.85	1769.47	1468.29	839.31	1541.37	1468.34	1108.51
黑龙江	4348.45	1681.88	1185.96	723.58	1363.62	1190.87	1082.96
海南	5673.65	780.10	1342.29	729.86	1830.80	1141.81	783.34
四川	5571.69	1483.54	1226.14	1020.16	1757.52	1369.47	735.26
贵州	4565.85	1209.88	1102.99	857.55	1395.28	1331.43	578.33
甘肃	4182.47	1470.26	1139.85	660.48	1289.80	1158.30	874.05
青海	4260.27	1394.28	1055.15	723.23	1293.45	967.90	854.25

资料来源：《中国统计年鉴》2012。

例 1 是一个二维的 4×5 列联表，位于列的属性变量眼睛的颜色有 4 个水平：深色、棕色、蓝色、浅色；位于行的属性变量头发的颜色有 5 个水平：金色、红色、棕色、深色、黑色。

在实际问题中，为了克服数量级对我们分析问题的影响，通常把频数的矩阵变换成概率矩阵 $P=(p_{ij})=(n_{ij}/n..)$，则例 1 对应的概率矩阵为：

眼睛颜色	头发颜色					合计
	金色	红色	棕色	深色	黑色	
深色	0.018192	0.00891	0.07481	0.126415	0.015779	0.2441
棕色	0.063672	0.015593	0.16874	0.07648	0.004826	0.3293
蓝色	0.060516	0.007054	0.044737	0.02042	0.000557	0.1333
浅色	0.127715	0.021533	0.108409	0.034899	0.000743	0.2933
合计	0.270095	0.053091	0.396696	0.258214	0.021905	1

例 2 中，我们有 10 个样品、7 个指标变量，在分析问题时，同样的，如果指标（变量）的量纲不同以及数量级相差很大时，我们要进行去量纲的处理，可以令每个数据除于总数：$P = (p_{ij}) = (x_{ij}/x_{..})$，则例 2 对应的权重矩阵（为叙述方便起见，我们也称它为概率矩阵）：

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
山西	0.029156	0.011979	0.010880	0.006824	0.012190	0.011631	0.006976
内蒙古	0.040664	0.020601	0.011625	0.009529	0.016418	0.014849	0.010156
辽宁	0.043061	0.015198	0.011354	0.007616	0.015562	0.013230	0.009901
吉林	0.034850	0.014500	0.012032	0.006878	0.012631	0.012032	0.009084
黑龙江	0.035633	0.013782	0.009718	0.005929	0.011174	0.009758	0.008874
海南	0.046492	0.006392	0.010999	0.005981	0.015002	0.009356	0.006419
四川	0.045657	0.012157	0.010047	0.008360	0.014402	0.011222	0.006025
贵州	0.037414	0.009914	0.009038	0.007027	0.011433	0.010910	0.004739
甘肃	0.034273	0.012048	0.009340	0.005412	0.010569	0.009492	0.007162
青海	0.034910	0.011425	0.008646	0.005926	0.010599	0.007931	0.007000

例 1，例 2 都是分析两个变量的对应关系：眼睛颜色与头发的颜色，省份与消费支出；并且两者的概率矩阵从形式看是相同的，因此，我们分析这两种类型的数据时，把原始数据转换成概率矩阵，然后只需从概率矩阵从发探讨 R 型因子分析（列数据）及 Q 型因子分析（行数据）。

则我们把样本矩阵统一为 $P = (p_{ij})$：

p_{11} p_{21} $\vdots$ p_{n1}	p_{12} p_{22} $\vdots$ p_{n2}	$\cdots$ $\cdots$ $\vdots$ $\cdots$	p_{1p} p_{2p} $\vdots$ p_{np}	$p_{1\cdot}$ $p_{2\cdot}$ $\vdots$ $p_{n\cdot}$
$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot p}$	1

这里，(1) 对列联表：$p_{ij} = \dfrac{n_{ij}}{n_{..}}, p_{i\cdot} = \sum_{j=1}^{p} p_{ij} = \sum_{j=1}^{p} \dfrac{n_{ij}}{n_{..}}, p_{\cdot j} = \sum_{i=1}^{n} p_{ij} = \sum_{i=1}^{n} \dfrac{n_{ij}}{n_{..}}$

(2) 对定量指标变量：$p_{ij} = \dfrac{x_{ij}}{x_{..}}, p_{i\cdot} = \sum_{j=1}^{p} p_{ij} = \sum_{j=1}^{p} \dfrac{x_{ij}}{x_{..}}, p_{\cdot j} = \sum_{i=1}^{n} p_{ij} = \sum_{i=1}^{n} \dfrac{x_{ij}}{x_{..}}$

显然有 $\sum_{i=1}^{n} p_{i\cdot} = \sum_{j=1}^{p} p_{\cdot j} = 1$。

9.1.2 对应分析的基本思想

由于 R 型因子分析和 Q 型因子分析都是反映一个整体的不同侧面，因此它们之间一定存在内在的联系。对应分析就是通过一个过渡矩阵 Z 将二者有机地结合起来，具体地说，首先给出变量（列）的协差阵 $A = Z'Z$ 和样品（行）的协差阵 $B = ZZ'$，由于 $Z'Z$ 和 ZZ' 有相同的非零特征根记为

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m, 0 < m \leqslant \min(p, n)$$

如果 A 的特征根 λ_i 对应的特征向量为 U_i，则 B 的特征根 λ_i 对应的特征向量就是 $ZU_i \triangleq V_i$，根据这个结论就可以很方便的借助 R 型因子分析而得到 Q 型因子分析的结果。因为求出 A 的特征根和特征向量后很容易地写出变量点协差阵对应的因子载荷阵，记为 F，则

$$F=\begin{bmatrix} u_{11}\sqrt{\lambda_1} & u_{12}\sqrt{\lambda_2} & \cdots & u_{1m}\sqrt{\lambda_m} \\ u_{21}\sqrt{\lambda_1} & u_{22}\sqrt{\lambda_2} & \cdots & u_{2m}\sqrt{\lambda_m} \\ \vdots & \vdots & & \vdots \\ u_{p1}\sqrt{\lambda_1} & u_{p2}\sqrt{\lambda_2} & \cdots & u_{pm}\sqrt{\lambda_m} \end{bmatrix}$$

这样一来样品点协差阵 B 对应的因子载荷阵记为 G，则

$$G=\begin{bmatrix} v_{11}\sqrt{\lambda_1} & v_{12}\sqrt{\lambda_2} & \cdots & v_{1m}\sqrt{\lambda_m} \\ v_{21}\sqrt{\lambda_1} & v_{22}\sqrt{\lambda_2} & \cdots & v_{2m}\sqrt{\lambda_m} \\ \vdots & \vdots & & \vdots \\ v_{n1}\sqrt{\lambda_1} & v_{n2}\sqrt{\lambda_2} & \cdots & v_{nm}\sqrt{\lambda_m} \end{bmatrix}$$

由于 A 和 B 具有相同的非零特征根，而这些特征根又正是各个公共因子的方差，因此可以用相同的因子轴同时表示变量点和样品点，即把变量点和样品点同时反映在具有相同坐标轴的因子平面上，以便对变量点和样品点一起考虑进行分类。

鉴于对应分析在列联表中的广泛应用性，本章主要从列联表出发探讨相关的问题。

9.2　列联表及列联表分析简介

列联表是由两个以上的属性变量进行交叉分类的频数分表。一般，若总体中的个体可按两个属性 A 与 B 分类，A 有 p 个水平 $A_1,A_2,\cdots,A_p$，B 有 n 个水平 $B_1,B_2,\cdots,B_n$，从总体中抽取大小为 $n..$ 的样本，设其中有 n_{ij} 个个体的属性属于水平 A_i 和 B_j，n_{ij} 称为频数，将 $n\times p$ 个 n_{ij} 排列为一个 n 行 p 列的二维列联表，简称 $n\times p$ 表。

$n\times p$ 列联表

行＼列	A_1	A_2	$\cdots$	A_p	合计
B_1	n_{11}	n_{12}	$\cdots$	n_{1p}	n_1
B_2	n_{21}	n_{22}	$\cdots$	n_{2p}	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
B_n	n_{n1}	n_{n2}	$\cdots$	n_{pp}	n_n
合计	n_1	n_2	$\cdots$	n_p	n

这里，行频数总和 $n_{i.}=\sum\limits_j n_{ij}$，列频数 $n._j=\sum\limits_i n_{ij}$，频数总和 $n..=\sum\limits_{ij} n_{ij}$。用 p_{ij} 表示第 ij 个格子频数占总频数的理论比例（概率），显然，$p_{ij}=\dfrac{E(n_{ij})}{n..}$，这里 $E(n_{ij})$ 为对 n_{ij} 的数学期望，而相应的第 i 行的理论比例（概率）$p_{i.}$ 及第 j 列的理论比例 $p._j$ 分别为：$p_{i.}=\sum\limits_{j=1}^{p} p_{ij}$ 和 $p._j=\sum\limits_{i=1}^{n} p_{ij}$。

若所考虑的属性多于两个，也可按类似的方式作出列联表，称为多维列联表。本节只考察二维列联表的相关问题。

列联表分析的基本问题是，判明所考察的各属性之间有无关联，即是否独立。

例 3 某项研究欲研讨患肺癌与吸烟是否有关，调查了 106 个志愿者，数据如下：

问题表述为：

	吸烟(人)	不吸烟(人)	合计(人)
患肺癌	60	3	63
未患肺癌	32	11	43
合计	92	14	106

H_0：患肺癌与否和吸烟与否相互独立　　H_1：患肺癌与否和吸烟与否不相互独立

也就是说要检验行和列的独立性，当行列变量独立时，一个观察值分配到第 ij 个格子的理论概率 p_{ij} 应该等于行列两个概率之积 $p_{ij}=p_{i\cdot}p_{\cdot j}$，则零假设为：

$$H_0: p_{ij}=p_{i\cdot}p_{\cdot j}$$

这时，在零假设下，它的估计值 $\hat{p}_{ij}=\hat{p}_{i\cdot}\hat{p}_{\cdot j}=\frac{n_{i\cdot}}{n_{\cdot\cdot}}\frac{n_{\cdot j}}{n_{\cdot\cdot}}$，则每个单元格频数的期望值为：

$$E_{ij}\approx\hat{p}_{ij}\times n_{\cdot\cdot}=\frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}$$

如果期望频数和观测频数相差不大，则零假设可能是正确的；如果二者差别很大，则零假设可能不成立。

$$\text{检验统计量 }\chi^2=\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\approx\frac{\sum_{i=1}^{n}\sum_{j=1}^{p}\left(n_{ij}-\frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}}$$

在零假设成立时，该统计量近似服从自由度为 $(n-1)(p-1)$ 的 χ^2 分布。当该统计量的值很大(或 p 值很小)时，就可以拒绝零假设，认为这两个变量不相互独立。例 3 $\chi^2=9.6636$，自由度为 1，计算 $p=0.00188$，在显著性水平 $\alpha=0.05$ 下拒绝原假设，即说明吸烟与肺癌不独立，存在一定的相关性。

9.3 对应分析的基本理论

在上一节中，我们通过列联表的卡方检验探究列联表中变量间的联系。问题在于：当属性变量 A 和 B 的水平较多时，很难透过列联表直观地揭示出变量之间的联系以及变量各水平之间的联系。主要表现在：

首先，由于变量的水平数较多使得交叉列联表行列数剧增，列联表庞大，不易于对列联表的直观观察。更主要的是，由于列联表的单元格数较多，极不易于揭示列联表中行列变量之间的联系。

其次，在变量水平数较多但样本量却不足够大时，生产的交叉列联表中会出现数据“稀疏”现象，不易于卡方检验等分析方法的运用。

怎样简化列联表的结构?可以利用降维的思想，如因子分析和主成分分析。但因子分析

的缺陷是在于无法同时进行 R 型因子分析和 Q 型因子分析。下面我们从二维列联表从发探讨两个属性变量的对应分析。

9.3.1　距离与总惯量

为了对列联表进行对应分析，首先，我们把频数的矩阵变换成概率矩阵 $P = (p_{ij}) = (n_{ij}/n..)$，则对应矩阵为：

概率矩阵 P

行＼列	A_1	A_2	…	A_p	合计
B_1	p_{11}	p_{12}	…	p_{1p}	p_1
B_2	p_{21}	p_{22}	…	p_{2p}	p_2
⋮	⋮	⋮	⋮	⋮	⋮
B_n	p_{n1}	n_{n2}	…	n_{pp}	n_n
合计	p_1	p_2	…	p_p	1

这里，$p_{ij} = \dfrac{n_{ij}}{n}, p_{i\cdot} = \sum\limits_{j=1}^{p} p_{ij} = \sum\limits_{j=1}^{p} \dfrac{n_{ij}}{n}, p_{\cdot j} = \sum\limits_{i=1}^{n} p_{ij} = \sum\limits_{i=1}^{n} \dfrac{n_{ij}}{n}$

显然有 $\sum\limits_{i=1}^{n} p_{i\cdot} = \sum\limits_{j=1}^{p} p_{\cdot j} = 1$。

如果我们将行数据看成 p 维空间的点，则其 n 个点的坐标用 $p_i^r = \left[\dfrac{p_{i1}}{p_{i.}}, \dfrac{p_{i2}}{p_{i.}}, \cdots, \dfrac{p_{ip}}{p_{i.}}\right]'$，$i = 1, \cdots, n$ 表示。同样的，如果我们将列数据看成是 n 维空间的点，则其 p 个点的坐标用 $p_j^c = \left[\dfrac{p_{1j}}{p_{\cdot j}}, \dfrac{p_{2j}}{p_{\cdot j}}, \cdots, \dfrac{p_{nj}}{p_{\cdot j}}\right]', j = 1, \cdots, p$ 表示

为了简化列联表，我们引入距离的概念来分别描述 A 的各个状态（n 个样品）之间的接近程度。

变量 A 的第 k 个状态与第 l 个状态的普通欧氏距离：

$$d^2(k,l) = (p_k^r - p_l^r)'(p_k^r - p_l^r) = \sum_{j=1}^{p}\left(\frac{p_{kj}}{p_{k\cdot}} - \frac{p_{lj}}{p_{l\cdot}}\right)^2$$

如此定义的距离有一个缺点，就是受到变量 B 的各个状态的边缘概率的影响，如 B 的第 j 个状态出现的概率特别大时，在上述公式中，$\left(\dfrac{p_{kj}}{p_{k\cdot}} - \dfrac{p_{lj}}{p_{l\cdot}}\right)$ 部分的作用就被抬高了，因此我们用系数 $\dfrac{1}{p_{\cdot j}}$ 去乘距离公式，就得到一个加权的距离公式

$$D^2(k,l) = \sum_{j=1}^{p}\left(\frac{p_{kj}}{p_{k.}} - \frac{p_{lj}}{p_{l.}}\right)^2 / p_{\cdot j} = \sum_{j=1}^{p}\left(\frac{p_{kj}}{p_{k.}\sqrt{p_{\cdot j}}} - \frac{p_{lj}}{p_{l.}\sqrt{p_{\cdot j}}}\right)^2$$

也可以说上式是坐标为

$$\left[\frac{p_{i1}}{\sqrt{p_{\cdot 1}}\,p_{i.}}, \frac{p_{i2}}{\sqrt{p_{\cdot 2}}\,p_{i.}}, \cdots, \frac{p_{ip}}{\sqrt{p_{\cdot p}}\,p_{i.}}\right] \qquad i = 1, \cdots, n$$

的 A 变量的第 k 个状态（样品）与第 l 个状态（样品）的距离。故 R 型因子分析的概率矩阵变为：

$$P^r = \begin{pmatrix} \frac{p_{11}}{\sqrt{p_{\cdot 1}}\, p_{1\cdot}} & \frac{p_{12}}{\sqrt{p_{\cdot 2}}\, p_{1\cdot}} & \cdots & \frac{p_{1p}}{\sqrt{p_{\cdot p}}\, p_{1\cdot}} \\ \frac{p_{21}}{\sqrt{p_{\cdot 1}}\, p_{2\cdot}} & \frac{p_{22}}{\sqrt{p_{\cdot 2}}\, p_{2\cdot}} & \cdots & \frac{p_{2p}}{\sqrt{p_{\cdot p}}\, p_{2\cdot}} \\ \vdots & \vdots & & \vdots \\ \frac{p_{n1}}{\sqrt{p_{\cdot 1}}\, p_{n\cdot}} & \frac{p_{n2}}{\sqrt{p_{\cdot 2}}\, p_{n\cdot}} & \cdots & \frac{p_{np}}{\sqrt{p_{\cdot p}}\, p_{n\cdot}} \end{pmatrix}$$

因为第 j 列的平均值按照概率加权：

$$\sum_{i=1}^{n} \frac{p_{ij}}{\sqrt{p_{.j}} \cdot p_{i.}} p_{i.} = \frac{1}{\sqrt{p_{.j}}} \sum_{i=1}^{n} p_{ij} = \frac{p_{.j}}{\sqrt{p_{.j}}} = \sqrt{p_{.j}}$$

故各列的加权平均值：$p_J^{1/2} = (\sqrt{p_{\cdot 1}} \quad \sqrt{p_{\cdot 2}} \quad \cdots \quad \sqrt{p_{\cdot p}})'$，我们定义其为 n 个点的中心。

定义行的总惯量：n 个点与其重心的欧氏距离之和，记为 $I_I = \sum_{i=1}^{n} D^2(p_i^r, p_J^{1/2})$。

又 $I_I = \sum_{i=1}^{n} D^2(p_i^r, p_J^{1/2}) = \sum_{i=1}^{n}\sum_{j=1}^{p} p_{i\cdot}\left(\frac{p_{ij}}{p_{i\cdot}\sqrt{p_{\cdot j}}} - \sqrt{p_{\cdot j}}\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{(p_{ij} - p_{i\cdot}p_{\cdot j})^2}{p_{i\cdot}p_{\cdot j}}$
$= \frac{1}{n}\chi^2$

从上式可以看出，总惯量不仅仅反映行数据的各点与其重心的加权距离的总和，同时与 χ^2 统计量仅相差一个常数，由列联表的知识我们知道，χ^2 统计量反映了列联表的横栏与纵栏的相关关系，因此，此处总惯量也反映了两个属性变量各状态之间的相关关系，对应分析就是在对总惯量信息损失最小的前提下，简化数据结构以反映两属性变量之间的相关关系，在 *Spss* 软件进行对应分析中，系统会给出对总惯量信息的提取情况。

类似地，B 属性的两个状态 i 与 j 之间的加权距离为：

$$D^2(i,j) = \sum_{k=1}^{n}\left(\frac{p_{ki}}{\sqrt{p_{k\cdot}}\, p_{\cdot i}} - \frac{p_{kj}}{\sqrt{p_{k\cdot}}\, p_{\cdot j}}\right)^2$$

也可以说上式是坐标为

$$\left[\frac{p_{1k}}{\sqrt{p_{1\cdot}}\, p_{k\cdot}}, \frac{p_{2k}}{\sqrt{p_{2\cdot}}\, p_{k\cdot}}, \cdots, \frac{p_{nk}}{\sqrt{p_{n\cdot}}\, p_{k\cdot}}\right]' \qquad k = 1, \cdots, p$$

的 B 属性的第 i 个状态与第 j 个状态的距离。故 Q 型因子分析的概率矩阵变为：

$$P^c = \begin{pmatrix} \frac{p_{11}}{\sqrt{p_{1.}}\, p_{.1}} & \frac{p_{12}}{\sqrt{p_{1.}}\, p_{.2}} & \cdots & \frac{p_{1p}}{\sqrt{p_{1.}}\, p_{.p}} \\ \frac{p_{21}}{\sqrt{p_{2.}}\, p_{.1}} & \frac{p_{22}}{\sqrt{p_{2.}}\, p_{.2}} & \cdots & \frac{p_{2p}}{\sqrt{p_{2.}}\, p_{.p}} \\ \vdots & \vdots & & \vdots \\ \frac{p_{n1}}{\sqrt{p_{n.}}\, p_{.1}} & \frac{p_{n2}}{\sqrt{p_{n.}}\, p_{.2}} & \cdots & \frac{p_{np}}{\sqrt{p_{n.}}\, p_{.p}} \end{pmatrix}$$

$p_I^{1/2} = (\sqrt{p_{1\cdot}}, \quad \sqrt{p_{2\cdot}}, \quad \cdots, \quad \sqrt{p_{n\cdot}})'$

列数据的总惯量：$I_I = I_J = \frac{1}{n}\chi^2$。

9.3.2 R 型与 Q 型因子分析的对等关系

经过以上的数据变换，在引入加权距离函数之后，就可以直接计算属性变量各状态之间的距离，通过距离的大小来反映不同状态之间的接近程度，同类型的状态之间距离应当较短，而不同类型的状态之间的距离应当较长，据此可以对各种状态进行分类以简化数据结构。但是，这样做不能对两个属性变量同时进行分析，因此不计算距离，代之求协方差矩阵，进行因子分析，提取主因子，用主因子所定义的坐标轴作为参照系，对两个变量的各个状态进行分析。

先进行 R 型因子分析，将 P^r 矩阵的 p 列看作 p 个变量，计算 p 个变量的协方差矩阵 A，由前面可知各列的加权平均值：$p_J^{1/2} = (\sqrt{p_{\cdot 1}},\quad \sqrt{p_{\cdot 2}},\quad \cdots,\quad \sqrt{p_{\cdot p}})'$，则我们可得：

$A = (a_{ij})$

其中
$$a_{ij} = \sum_{a=1}^{n}\left(\frac{p_{ai}}{\sqrt{p_{\cdot i}p_{a\cdot}}} - \sqrt{p_{\cdot i}}\right)\left(\frac{p_{aj}}{\sqrt{p_{\cdot j}p_{a\cdot}}} - \sqrt{p_{\cdot j}}\right)p_{a\cdot}$$
$$= \sum_{a=1}^{n}\left(\frac{p_{ai}}{\sqrt{p_{\cdot i}}\sqrt{p_{a\cdot}}} - \sqrt{p_{\cdot i}}\sqrt{p_{a\cdot}}\right)\cdot\left(\frac{p_{aj}}{\sqrt{p_{\cdot j}}\sqrt{p_{a\cdot}}} - \sqrt{p_{\cdot j}}\sqrt{p_{a\cdot}}\right)$$
$$= \sum_{a=1}^{n}\left(\frac{p_{ai} - p_{\cdot i}p_{a\cdot}}{\sqrt{p_{\cdot i}p_{a\cdot}}}\right)\left(\frac{p_{aj} - p_{\cdot j}p_{a\cdot}}{\sqrt{p_{\cdot j}p_{a\cdot}}}\right)$$
$$\triangleq \sum_{a=1}^{n} z_{ai}z_{aj}$$

其中 $z_{ai} = \dfrac{p_{ai} - p_{\cdot i}p_{a\cdot}}{\sqrt{p_{\cdot i}p_{a\cdot}}}\quad a = 1,\cdots,n\qquad i = 1,\cdots,p$

令 $Z = (z_{ij})$，则有

$$A = Z'Z$$

类似上面的方法，进行 Q 型因子分析，将 P^c 矩阵的 n 行看作 n 个变量，计算 n 个变量的协方差矩阵 B，由前面可知各行的加权平均值：$p_I^{1/2} = (\sqrt{p_{1\cdot}},\quad \sqrt{p_{2\cdot}},\quad \cdots,\quad \sqrt{p_{n\cdot}})'$，则可求

$$B = (b_{KL})$$

其中

$$b_{KL} = \sum_{i=1}^{p}\left(\frac{p_{Ki}}{\sqrt{p_{K\cdot}p_{\cdot i}}} - \sqrt{p_{K\cdot}}\right)\left(\frac{p_{Li}}{\sqrt{p_{L\cdot}p_{\cdot i}}} - \sqrt{p_{L\cdot}}\right)p_{\cdot i}$$
$$= \sum_{i=1}^{p}\left(\frac{p_{Ki}}{\sqrt{p_{K\cdot}}\sqrt{p_{\cdot i}}} - \sqrt{p_{K\cdot}}\sqrt{p_{\cdot i}}\right)\left(\frac{p_{Li}}{\sqrt{p_{\cdot i}}\sqrt{p_{L\cdot}}} - \sqrt{p_{L\cdot}}\sqrt{p_{\cdot i}}\right)$$
$$= \sum_{i=1}^{p}\left(\frac{p_{Ki} - p_{\cdot i}p_{K\cdot}}{\sqrt{p_{\cdot i}p_{K\cdot}}}\right)\left(\frac{p_{Li} - p_{\cdot i}p_{L\cdot}}{\sqrt{p_{\cdot i}p_{L\cdot}}}\right) = \sum_{i=1}^{p} z_{Ki}z_{Li}$$

其中

$$z_{Ki} = \frac{p_{K\cdot} - p_{\cdot i}p_{K\cdot}}{\sqrt{p_{\cdot i}p_{K\cdot}}},\ z_{Li} = \frac{p_{L\cdot} - p_{\cdot i}p_{L\cdot}}{\sqrt{p_{\cdot i}p_{L\cdot}}}$$

从而

$$B = ZZ'$$

综上所述，若 R 型与 Q 型因子分析的协差阵分别为：$A = Z'Z$ 和 $B = ZZ'$。A 与 B 两矩阵明显存在着简单的对应关系。

为了进一步研究 R 型与 Q 型因子分析的对应关系，我们可以借助下面线性代数中的定理。

定理 9.1　A 与 B 的非零特征根相同，记为 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m$。

推论 9.1　如果 U 是 $Z'Z$ 的特征向量，则 ZU 是 ZZ' 的特征向量。如果 V 是 ZZ' 的特征向量，则 $Z'V$ 是 $Z'Z$ 的特征向量。

这是显然的，比如 U 是 $Z'Z$ 的特征向量，则有

$$Z'ZU = \lambda U \quad 两边左乘 Z 得：Z'Z(ZU) = \lambda(ZU)$$

即 ZU 是 $Z'Z$ 的特征向量。

这个定理为我们建立了因子分析中 R 型与 Q 型的关系。因此借助这个定理，我们可以从 R 型因子分析出发而直接获得 Q 型因子分析的结果。又由于 A 与 B 有相同的非零特征根，而这些特征根正是各个公因子所解释的方差，或提取的总惯量的份额，既有 $\sum_{i=1}^{m}\lambda_i = I_I = I_J$。那么在变量 B 的 p 维空间 R^p 中的第一主因子、第二主因子、…、直到第 r 个主因子与变量 A 的 n 维空间 R^n 中相对应的各个主因子在总方差中所占的百分比就完全相同。这样就可以用相同的因子轴去同时表示两个属性变量的各个水平，把两个变量的各个水平同时反映在具有相同坐标轴的因子平面上，以直观的反映两个属性变量及各个水平之间的相关关系，一般的情况下，我们取两个公因子，这样就可以在一张二维的图上同时画出两个变量的各个状态。

9.4　对应分析的步骤

第一步　由原始资料阵 X 或列联表出发，计算规格化的概率矩阵 $P = (p_{ij})$

第二步　计算过渡矩阵 $Z = (z_{ij})$，其中

$$z_{ij} = \frac{p_{ij} - p_{i\cdot}\, p_{\cdot j}}{\sqrt{p_{i\cdot}\, p_{\cdot j}}}$$

第三步　进行因子分析

(1)R 型因子分析

计算协差阵 $A = Z'Z$ 的特征根 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$，按其累计百分比 $\sum_{a=1}^{m}\lambda_m / \sum_{a=1}^{p}\lambda_a \geqslant 85\%$，取前 m 个特征根 $\lambda_1, \lambda_2, \cdots, \lambda_m$，并计算相应的单位特征向量记为：$u_1, u_2, \cdots, u_m$，从而得到因子载荷阵：

$$F = \begin{pmatrix} u_{11}\sqrt{\lambda_1} & u_{12}\sqrt{\lambda_2} & \cdots & u_{1m}\sqrt{\lambda_m} \\ u_{21}\sqrt{\lambda_1} & u_{22}\sqrt{\lambda_2} & \cdots & u_{2m}\sqrt{\lambda_m} \\ \vdots & \vdots & & \vdots \\ u_{p1}\sqrt{\lambda_1} & u_{p2}\sqrt{\lambda_2} & \cdots & u_{pm}\sqrt{\lambda_m} \end{pmatrix}$$

(2)Q 型因子分析

协差阵 $B = ZZ'$，有定理 9.1 可知，B 的特征根跟 A 相同，因此，这里也取 m 个特征根 λ_1，$\lambda_2, \cdots, \lambda_m$，由推 9.1 可知，其对应于矩阵 $B = ZZ'$ 的单位特征向量 $v_1 = Zu_1, v_2 = Zu_2, \cdots, v_m = Zu_m$，从而得到 Q 型的因子载荷阵为：

$$G = \begin{pmatrix} v_{11}\sqrt{\lambda_1} & v_{12}\sqrt{\lambda_2} & \cdots & v_{1m}\sqrt{\lambda_m} \\ v_{21}\sqrt{\lambda_1} & v_{22}\sqrt{\lambda_2} & \cdots & v_{2m}\sqrt{\lambda_m} \\ \vdots & \vdots & & \vdots \\ v_{p1}\sqrt{\lambda_1} & v_{p2}\sqrt{\lambda_2} & \cdots & v_{pm}\sqrt{\lambda_m} \end{pmatrix}$$

(3) 若取 $m = 2$，则在 F,G 前两列在二维表上作出对应图。

本章思考与练习

1. 简述对应分析的基本思想及特点。
2. 试述对应分析中总惯量的意义。

第十章　多维标度分析

在实际中往往碰到这样的问题：有 n 个由多项指标（变量）反映的对象，但反映对象的指标个数是多少不清楚，甚至指标本身是什么也是模糊的，更谈不上直接测量或观察它，仅仅所能知道的是这 n 个对象之间的某种距离（不一定是通常的欧氏距离）或者某种相似性，我们希望仅由这种距离或者相似性出发，在较低维的欧氏空间把 n 个对象（作为几何点）的图形描绘出来，从而尽可能揭示这 n 个对象之间的真实结构关系。这就是多维标度法所要研究的问题。

一个经典的例子是利用城市之间的距离来绘制地图。

例 1　表 1 是某国十城市之间的飞行距离，我们如何在平面坐标上据此标出这 10 城市之间的相对位置，使之尽可能接近表中的距离数据呢？

表 1　10 城市间的飞行距离

城市	1	2	3	4	5	6	7	8	9	10
1	0	245	223	251	200	314	217	473	170	215
2	247	0	355	81	169	448	226	584	245	53
3	222	353	0	371	348	97	154	254	117	301
4	252	78	373	0	244	465	241	598	260	75
5	197	164	351	245	0	445	284	603	273	171
6	312	446	99	467	445	0	247	169	213	394
7	219	226	155	237	281	248	0	385	56	169
8	474	587	254	600	598	173	384	0	352	535
9	170	242	118	257	273	215	55	350	0	190
10	214	56	298	75	172	395	173	536	190	0

上述的问题可以表述为：已知 10 个城市两两之间的距离矩阵 $D = (d_{ij})_{10\times10}$，我们的目的是求 R^2 中的 10 个点 $X_1, X_2, \cdots, X_{10}$，使得：

$$\hat{d}_{ij}^2 = (X_i - X_j)'(X_i - X_j) \quad i,j = 1,2,\cdots,10$$

尽可能的接近 d_{ij}^2。其中，$X_i = (X_{i1}, X_{i2})'$ 就是我们要求的第 i 个城市的相对坐标点。

多维标度法（Multi-Dimensional Scaling）就是解决类似问题的一种方法，它是一种在低维空间展示“距离”数据结构的多元数据分析技术，简称 MDS。

多维标度法内容丰富、方法较多，按相似性（距离）矩阵的个数和 MDS 模型的性质，MDS 可分为：古典多维标度 CMDS（一个矩阵，无权重模型）、重复多维标度 Replicated MDS（几个矩阵，无权重模型）、权重多维标度 WMDS（几个矩阵，权重模型）。本章仅介绍常用的古典多维标度法。

10.1 距离阵和经典解

10.1.1 欧式距离阵

在解决问题之前，我们首先明确与多维标度法相关的数据概念。

我们这里研究的距离不限于通常的欧氏距离，首先，对距离的意义加以拓广，给出如下的距离阵的定义。

定义 10.1 一个 $n \times n$ 阶的矩阵 $D = (d_{ij})_{n \times n}$，如果满足条件：

(1)$D = D'$

(2)$d_{ij} \geqslant 0,\ d_{ii} = 0,\quad i,j = 1,2,\cdots,n$

则矩阵 D 为广义距离阵，d_{ij} 称为第 i 点与第 j 点间的距离。

从例 1 的讨论可知，多维标度分析要解决的问题是从 n 个对象已知的距离矩阵 $D = (d_{ij})$ 出发，求正整数 r 和 R^r 中的 n 个点 $X_1, X_2, \cdots, X_n$，使得

$$\hat{d}_{ij}^2 = (X_i - X_j)'(X_i - X_j) \qquad i,j = 1,2,\cdots,n$$

在某种意义下尽可能的接近 d_{ij}^2，若$\hat{d}_{ij}^2 = d_{ij}^2$，我们称具有这种性质的距离矩阵 $D = (d_{ij})$ 为欧氏距离阵。

定义 10.2 对于一个 $n \times n$ 的距离阵 $D = (d_{ij})_{n \times n}$，如果存在某个正整数 r 和 R^r 中的 n 个点 $X_1, X_2, \cdots, X_n$，使得

$$d_{ij}^2 = (X_i - X_j)'(X_i - X_j) \qquad i,j = 1,2,\cdots,n$$

则称 D 为欧氏距离阵。

令 $X = (X_1, X_2, \cdots, X_n)'$，则称 X 为 D 的一个解，在多维标度法中，我们称 X 为距离阵 D 的一个拟合构图。所谓的拟合构图，其意义是有了这 n 个点的坐标，可以在 R^r(r 一般取 1，2，3) 画出图来，使得它们的距离阵$\hat{D}$ 和原始的对象距离阵 D 接近，给出原始 n 个对象关系一个有意义的解释，特别地，如果 $\hat{D} = D$，则称 X 为 D 的一个构图，也就是说，如果 D 是欧氏距离阵，那么相应的 X 是 D 的一个构图。

10.1.2 欧式距离阵的判定定理

上一节，我们给出欧氏距离阵的定义，在这一节中，我们要讨论如何判别一个距离阵 D 是欧氏距离阵。

为了便于理解，我们从实际的例子出发，假设 n 个城市的已知距离阵为 $D = (d_{ij})$，我们目的是求 R^r 欧氏空间的 n 个点，第 i 个城市对应的点记为 X_i，则 X_i 的坐标记作 $X_i = (X_{i1}, X_{i2}, \cdots, X_{ir})'$。

令

$$B = (b_{ij})_{n \times n},$$

其中：

$$b_{ij} = \frac{1}{2}\left(-d_{ij}^2 + \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2 + \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2\right) \tag{1}$$

d_{ij} 为 i 城市与 j 城市之间的距离。

定理 10.1 一个 $n\times n$ 的距离阵 D 是欧氏距离阵的充要条件是 $B\geqslant 0$。

证明:必要性,设 $D=(d_{ij})$ 是欧氏距离阵,根据定义(2),存在 $X_1,X_2,\cdots,X_n\in\mathbf{R}^r$,使得

$$\begin{aligned}d_{ij}^2&=(X_i-X_j)'(X_i-X_j)\\&=X_i'X_i+X_j'X_j-X_j'X_i-X_i'X_j\\&=X_i'X_i+X_j'X_j-2X_i'X_j\end{aligned}\tag{2}$$

则可得

$$\frac{1}{n}\sum_{i=1}^{n}d_{ij}^2=X_j'X_j+\frac{1}{n}\sum_{i=1}^{n}X_i'X_i-\frac{2}{n}\sum_{i=1}^{n}X_i'X_j\tag{3}$$

同理可得

$$\frac{1}{n}\sum_{j=1}^{n}d_{ij}^2=X_i'X_i+\frac{1}{n}\sum_{j=1}^{n}X_j'X_j-\frac{2}{n}\sum_{j=1}^{n}X_i'X_j\tag{4}$$

最后,$\frac{1}{n}\sum_{j=1}^{n}(\frac{1}{n}\sum_{i=1}^{n}d_{ij}^2)=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}d_{ij}^2$

$$=\frac{1}{n}\sum_{i=1}^{n}X_i'X_i+\frac{1}{n}\sum_{j=1}^{n}X_j'X_j-\frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}X_i'X_j\tag{5}$$

把(3)(4)(5)代入(1),

$$\begin{aligned}b_{ij}&=\frac{1}{2}\left(-d_{ij}^2+\frac{1}{n}\sum_{j=1}^{n}d_{ij}^2+\frac{1}{n}\sum_{i=1}^{n}d_{ij}^2-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}d_{ij}^2\right)\\&=\frac{1}{2}\left(2X_i'X_j-\frac{2}{n}\sum_{j=1}^{n}X_i'X_j-\frac{2}{n}\sum_{i=1}^{n}X_i'X_j+\frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}X_i'X_j\right)\\&=(X_i'X_j-X'_i\overline{X}-\overline{X}'X_j+\overline{X}'\overline{X})\\&=(X_i-\overline{X})'(X_j-\overline{X})\end{aligned}$$

其中:$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$,用矩阵表示为

$$B=(b_{ij})_{n\times n}=\begin{bmatrix}(X_1-\overline{X})'\\\vdots\\(X_n-\overline{X})'\end{bmatrix}(X_1-\overline{X},\cdots,X_n-\overline{X})\geqslant 0$$

必要性得证。

充分性,若 $B\geqslant 0$,那么 D 是欧氏型的,并且按以下方法构造的 X 正好为 D 的一个构图。

记 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_r$ 为 B 的正特征根,$\lambda_1,\lambda_2,\cdots,\lambda_r$ 对应的单位特征向量

$$u_1,u_2,\cdots,u_r$$

$\Gamma=(u_1,u_2,\cdots,u_r)$ 是单位特征向量为列组成的矩阵,则令

$$X=(\sqrt{\lambda_1}u_1,\sqrt{\lambda_2}u_2,\cdots,\sqrt{\lambda_r}u_r)=(x_{ij})_{n\times r}\tag{6}$$

X 矩阵中每一行对应空间中的一个点,第 i 行即为 X_i'。

令 $\Lambda=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_r)$,那么 $X=\Gamma\Lambda^{1/2}$,可得

$$B=\Gamma\Lambda\Gamma'=XX'\tag{7}$$

即 $b_{ij}=X_i'X_j$,由(1)式 $b_{ij}=\frac{1}{2}\left(-d_{ij}^2+\frac{1}{n}\sum_{j=1}^{n}d_{ij}^2+\frac{1}{n}\sum_{i=1}^{n}d_{ij}^2-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}d_{ij}^2\right)$,我们可到

$(X_i-X_j)'(X_i-X_j)=X_i'X_i+X_j'X_j-2X_i'X_j=b_{ii}+b_{jj}-2b_{ij}=d_{ij}^2$

这样说明 X 正好为 D 的一个构图，D 是欧氏型的。

通过上面的讨论我们知道，只要按公式(1) 求出各个点对之间的内积，求得内积矩阵 B 的 r 个非零特征值及所对应的一组特征向量，据公式(6) 即可求

出 X 矩阵的 r 个列向量或空间 n 个点的坐标。

10.1.3 多维标度的经典解

这里需要特别注意，并非所有的距离阵都存在一个 r 维的欧氏空间和 n 个点，使得 n 个点之间的距离等于 D。因而，并不是所有的距离阵都是欧氏距离阵，还存在非欧氏距离阵。

当距离阵 D 为欧氏时，可求得一个 D 的构图 X，当距离阵不是欧氏时，只能求得 D 的拟合构图。在实际应用中，即使 D 为欧氏，一般也只求 $r=2$ 或 3 的低维拟合构图。

值得注意的是，由于多维标度法求解的 n 个点仅仅要求它们的相对欧氏距离与 D 相近，也就是说，只与相对位置相近而与绝对位置无关，根据欧氏距离在正交变换和平移变换下的不变性，显然所求得解并不唯一。

根据上述古典多维标度法的基本思想及方法，可给出求古典解的一般步骤：

(1) 根据距离阵数据，按照公式(1) 计算出 b_{ij}；

(2) 根据 b_{ij} 构造出矩阵 $B=(b_{ij})$；

(3) 计算内积矩阵 $B=(b_{ij})$ 的特征值 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$ 和 r 个最大特征值 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_r > 0$ 对应的单位特征向量。其中，r 的确定有两种方法：一是事先确定 $r=1,2$ 或 3；二是通过计算前 r 个大于零的特征值占全体特征值的比例 κ 确定。

$$\kappa = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_r}{|\lambda_1| + |\lambda_2| + \cdots + |\lambda_n|} \geqslant \kappa_0 \tag{8}$$

κ_0 预先给定的变差贡献比例。

(4) 根据(6) 式计算 $\hat{X}$，得到 r 维拟合构图(简称古典解)。这里需要注意，如果 λ_i 中有负值，表明 D 是非欧氏型的。

10.2 实　例

例 1　表 1 是某国十城市之间的飞行距离，我们如何在平面坐标上据此标出这 10 城市之间的相对位置，使之尽可能接近表中的距离数据呢？

表 1　10 城市间的飞行距离

城市	1	2	3	4	5	6	7	8	9	10
1	0	245	223	251	200	314	217	473	170	215
2	247	0	355	81	169	448	226	584	245	53
3	222	353	0	371	348	97	154	254	117	301
4	252	78	373	0	244	465	241	598	260	75
5	197	164	351	245	0	445	284	603	273	171
6	312	446	99	467	445	0	247	169	213	394
7	219	226	155	237	281	248	0	385	56	169
8	474	587	254	600	598	173	384	0	352	535

续表

城市	1	2	3	4	5	6	7	8	9	10
9	170	242	118	257	273	215	55	350	0	190
10	214	56	298	75	172	395	173	536	190	0

解:首先可求得内积矩阵 $\boldsymbol{B}$,结果见表 2。

表 2　10 城市的矩阵 B

22123.72	2940.82	−3647.23	6417.07	18654.77	−9206.33	−9183.08	−25750.1	−2527.83	178.22
2645.82	43979.92	−30768.6	45663.67	35400.87	−49233.7	−149.98	−73387	−7063.73	32912.82
−3654.53	−30487.4	20081.02	−32039.2	−23033.5	34250.92	1367.17	52720.12	3941.42	−23146
6176.42	45716.02	−32542.5	53722.27	24691.47	−52216.1	1125.62	−76882.9	−6073.13	36282.92
19604.07	36390.17	−23498.4	24789.92	55539.62	−42036	−9081.73	−78805.3	−8457.48	25555.07
−8795.48	−48751.9	34069.57	−53374.1	−42604.9	57844.47	1609.72	89586.67	6990.47	−36574.5
−9740.58	−468.48	1320.97	1949.27	−8709.53	1455.87	6477.62	4118.07	2470.37	1126.42
−26135	−75251.9	53038.52	−78003.2	−76068	89206.42	4712.67	150193.6	14049.42	−55742.5
−2700.18	−6702.58	3881.37	−5480.83	−8983.63	6605.27	2475.02	14490.47	1548.27	−5133.18
475.77	32635.37	−21934.7	36355.12	25112.82	−36670.8	646.97	−56283.6	−4877.78	24540.77

B 的特征值为:$\lambda_1=372040$,$\lambda_2=47360$,$\lambda_3=16050$,$\lambda_4=5630$,$\lambda_5=-4020$,$\lambda_6=-2610$,$\lambda_7=1850$,$\lambda_8=290$,$\lambda_9=0$,$\lambda_{10}=-540$。

因此取 $r=2$。按照(8)式得到如下结果:

$$\kappa_2=\frac{\lambda_1+\lambda_2}{|\lambda_1|+|\lambda_2|+\cdots+|\lambda_{10}|}$$

$$=\frac{372040+47360}{372040+47360+16050+5630+4020+2610+1850+290+0+540}=0.961816$$

$\sqrt{\lambda_1}u_1$	$\sqrt{\lambda_2}u_2$	u_1	u_2
52.39478	112.359	0.0859	0.5163
201.0398	−46.0927	0.3296	−0.2118
−140.655	29.2486	−0.2306	0.1344
208.8472	−88.1375	0.3424	−0.405
193.7204	125.7429	0.3176	0.5778
−237.515	27.35528	−0.3894	0.1257
−10.6741	−61.3263	−0.0175	−0.2818
−383.781	−36.8872	−0.6292	−0.1695
−34.7062	−17.5405	−0.0569	−0.0806
151.3898	−44.7216	0.2482	−0.2055

10 个城市的坐标分别为:

(−52.39,112.359),(201.03,−46.09),(−140.66,29.24),(208.84,−88.14),(193.72,125.74),(−237.51,27.36),(−10.67,−61.33),(−383.78,−36.89),(−34.70,−17.54),(151.39,−44.72)。

计算结果表明,较大的特征值有两个,说明在二维平面上表示 10 城市间的相对位置是合适的。由于有特征值小于零,表明距离阵不是欧氏型,其结果为拟合构图。在此,城市是"对象",飞行里程是"相似性"。图 1 给出了 MDS 反映这 10 座城市相对位置的感知图。图中的 10

个点，每个点代表一个城市，相近的点代表飞行距离短的城市，相距较远的点代表飞行距离远的城市。(注：由于解的多样性，这里给出的图是 *Spss* 画出的图，但是相对位置不变的)

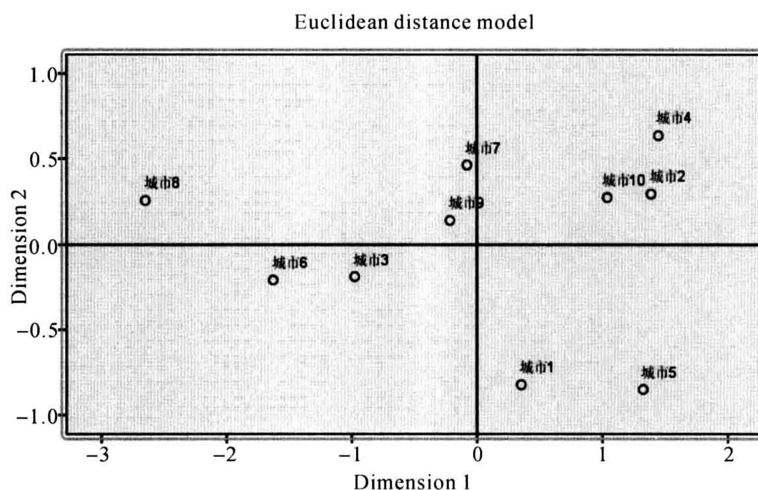


图 1

本章思考与练习

1. 简述多维标度分析的作业。
2. 试述多维标度分析的思想与方法。

第二篇　多元统计分析实验

多元统计分析是统计分析学中内容十分丰富，应用范围极为广泛的一个分支。在自然科学和社会科学的许多学科中，研究者都有可能需要分析处理有多个变量的数据的问题，能否从表面看起来杂乱无章的数据中发现和提炼出规律性的结论，不仅对所研究的专业领域要有很好的训练，而且要掌握必要的统计分析工具。

统计学科专业课程实验教学通过利用数据库、统计专业软件和多媒体技术，使教学内容更深入、更生动、更全面。统计实验教学不仅改变了传统的教学模式，而且教学观念也得到了更新和提高，把教师和学生从繁重枯燥的教学任务中解脱出来，使学生在轻松有趣的氛围中得到综合能力的提高，这也是一种寓教于乐。具体地说，统计实验教学以学生为主体，教师由主讲者的角色逐渐转变为学习活动的设计者和指导者，教学媒体不仅是教师的讲授工具，而且也是学生的认识工具；学生也从知识的“被动接受者”转变为积极参与教学、参与操作、发现知识、理解知识、掌握知识的“主动寻求者”。实践证明，实验教学有利于提高学生对统计学科专业课程的学习兴趣和教学效果。

多元分析的应用离不开计算机，一般的多元统计分析教材都没有上机实现的操作说明，为了使学生更好的掌握 SPSS 软件在各种多元统计方法中的应用，本书在实验部分配合实例概要介绍了 SPSS 软件的实际操作过程，这对于指导学生学习多元统计分析大有裨益。

本手册共编入与教材相结合的八个实验项目，每一项目中包含上机操作说明，上机操作图示及相关的数据分析。

实验一　均值向量和协方差阵的检验

1.1　实验背景

1999 年财政部、国家经贸委、人事部和国家计委联合发布了《国有资本金效绩评价规则》。其中，对竞争性工商企业的评价指标体系包括下面八大基本指标：

(1) 净资产收益率
(2) 总资产报酬率
(3) 总资产周转率
(4) 流动资产周转率
(5) 资产负债率
(6) 已获利息倍数
(7) 销售增长率
(8) 资本积累率

本实验选择来自三个行业的 35 家代表性上市公司年报中的这八项财务数据，在 *SPSS* 中利用均值向量和协方差阵的检验功能对三个行业间上市公司的运营情况进行分析，分析的问题主要包括：

(1) 不同行业的上市公司运营能力有无显著差异；
(2) 若有差异，差异来自哪些行业，作不同行业的运营能力的比较分析；
(3) 各行业(总体) 协方差阵相等的检验。

1.2　实验步骤和结果分析

(一) 实验数据

例 1.1 借助指标体系(净资产收益率、总资产报酬率、资产负债率、总资产周转率、流动资产周转率、已获利息倍数、销售增长率及资本积累率) 对我国上市公司的运营情况进行分析。表 1-1 所列的是 35 家上市公司 2000 年年报数据，其中，11 家上市公司来自于电力、煤气及水的生产和供应业，15 家来自房地行业，9 家来自信息技术业。

表 1-1

行业	公司简介	净资产收益率(%)	总资产报酬率(%)	资产负债率(%)	总资产周转率	流动资产周转率(%)	已获利息倍数	销售增长率(%)	资本积累率(%)
电力煤气及水的生产和供应业	深能源 A	16.85	12.35	42.32	0.37	1.78	7.18	45.73	54.54
	深南电 A	22	15.3	46.51	0.76	1.77	15.67	48.11	19.41
	富龙热力	8.97	7.98	30.56	0.17	0.58	10.43	17.8	9.44
	穗恒运 A	10.25	8.99	40.44	0.46	2.46	5.06	11.06	1.09
	粤电力 A	20.81	20	35.87	0.43	1.25	34.89	24.77	12.67
	韶能股份	8.86	7.52	27.59	0.24	0.84	20.59	−3.5	54.02
	惠天热电	10.98	7.94	49.3	0.36	0.69	12.43	16.88	3.52
	城投控股	8.85	8.88	36.2	0.13	0.41	8.53	−11.49	2.44
	大连热电	9.03	7.41	46.89	0.28	0.79	6.86	16.23	−1.52
	龙电股份	12.07	8.7	16.81	0.28	0.68	29.75	4.11	63.06
	华银电力	6.85	6.12	41.93	0.24	0.65	4.38	11.2	3.8
房地产行业	长春经开	9.85	10.5	31.23	0.34	0.4	17.13	18.05	7.18
	兴业房产	1.07	1.52	66.91	0.21	0.24	1.53	−31.93	1.08
	金丰投资	19.44	7.01	73.34	0.26	0.3	7.02	71.22	12.73
	新黄浦	7.61	5.92	39.64	0.16	0.17	4.2	14.77	7.91
	浦东金桥	4.24	3.99	37.3	0.2	0.25	3.98	−9.24	4.69
	外高桥	1.673	1.92	49.05	0.03	0.05	1.06	−21.74	0.24
	中华企业	8.78	6.28	57.42	0.17	0.19	3.58	75.29	2.93
	渝开发 A	0.2	2.24	63.4	0.09	0.15	1.07	−12.56	0.29
	辽房天	8.12	3.98	69.1	0.1	0.72	2.65	−35.83	3.16
	粤宏远 A	0.42	1.16	37.42	0.09	0.15	1.59	19.18	0.43
	ST 中福	5.17	6.62	65.48	0.16	0.21	1.33	−19.91	23.74
	倍特高新	0.72	2.76	65.39	0.3	0.42	1.24	8.4	0.7
	三木集团	5.99	4.53	65.17	0.74	0.88	4.14	75.36	0.87
	寰岛实业	0.42	0.2	24.03	0.02	0.03	−8.18	−71.33	0.42
	中关村	9.32	4.48	67.76	0.32	0.37	16.42	−29.42	4.09
信息技术业	中兴通讯	18.78	11.09	69.15	0.93	1.08	4.79	80.8	23.27
	长城电脑	14.94	9.48	45.53	1.14	1.85	9.51	34.47	35.93
	青鸟华光	9.788	8.7	36.67	0.28	0.39	13.11	28.36	7.87
	清华同方	15.91	9.08	34.19	0.85	1.19	15.61	98.92	95.66
	永鼎光缆	9.4	8.67	32.75	0.79	1.25	13.49	41.75	6.33
	宏图高科	14.57	7.96	65.86	0.76	0.94	3.95	54.45	15.71
	海星科技	4.06	3.35	36.49	0.48	0.6	4.64	−16.28	1.69
	方正科技	27.48	16.69	57.13	2.51	2.87	7.4	63.27	32.02
	复华实业	5.58	4.1	44.24	0.28	0.41	3.77	12.92	2.3

注:1. 该表中,除大连热电的数据为母公司数据外,其他数据均来自于合并会计报表;

2. 除辽房天及中兴通讯外,其他公司的净资产收益率均为加权后的数值;

3. 除净资产收益率指标为直接取自会计年报外,其他各指标均是经过各企业年报提供数字计算而得,各指标的计算公司如下:

a. 总资产报酬率 $= \dfrac{\text{利润总额} + \text{财务费用}}{(\text{年初总资产} + \text{年末总资产})/2} \times 100\%$

b. 资产负债率 $= \dfrac{\text{年末负债总额}}{\text{年末资产总额}} \times 100\%$

c. 总资产周转率 $= \dfrac{\text{主营业务收入}}{(\text{年初总资产} + \text{年末总资产})/2}$

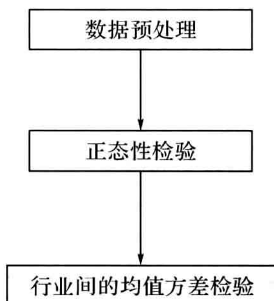
$$流动资产周转率 = \frac{主营业务收入}{(年初流动资产 + 年末流动资产)/2}$$

$$已获利息倍数 = \frac{利润总额 + 财务费用}{财务费用}$$

$$销售增长率 = \frac{本年主营业务收入 - 上年主营业务收入}{上年主营业务收入} \times 100\%$$

$$资本积累率 = \frac{年末股东权益 - 年初股东权益}{年初股东权益} \times 100\%$$

(二)实验步骤



(1) 数据预处理

第一步 Excel 处理:为了便于进行 SPSS 分析,将上述原始数据的 Excel 文档改为如下形式:

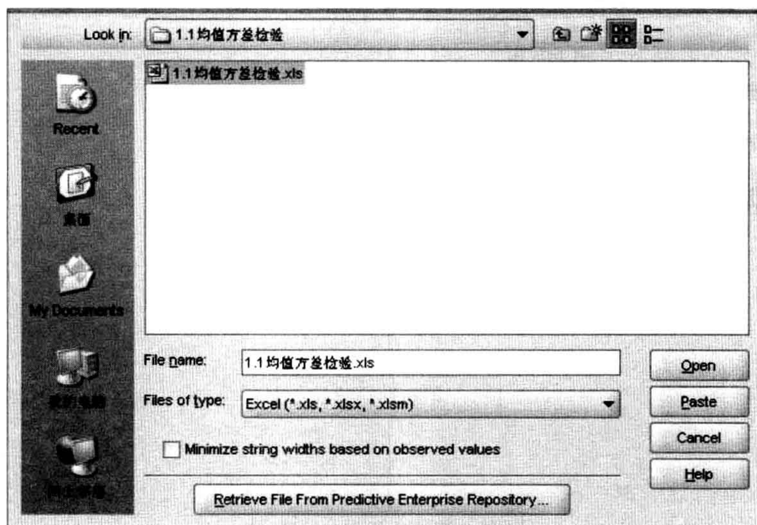
公司	行业	净资产收益率(%)	总资产报酬率(%)	资产负债率(%)	总资产周转率	流动资产周转率(%)	已获利息倍数	销售增长率(%)	资本积累率(%)
深能源 A	1	16.85	12.35	42.32	0.37	1.78	7.18	45.73	54.54
深南电 A	1	22	15.3	46.51	0.76	1.77	15.67	48.11	19.41
…	…	…	…	…	…	…	…	…	…
长春经开	2	9.85	10.5	31.23	0.34	0.4	17.13	18.05	7.18
兴业房产	2	1.07	1.52	66.91	0.21	0.24	1.53	−31.93	1.08
…	…	…	…	…	…	…	…	…	…
中兴通讯	3	18.78	11.09	69.15	0.93	1.08	4.79	80.8	23.27
长城电脑	3	14.94	9.48	45.53	1.14	1.85	9.51	34.47	35.93

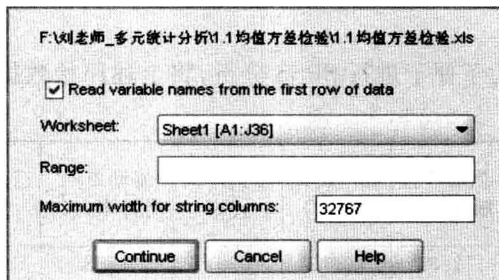
即分别用数字 1、2、3 来代替每个公司所在的行业,并将数据保存为“1.1 均值方差分析.xls”。

第二步数据导入:将 xls 格式的 Excel 数据导入到 SPSS 中:打开 SPSS → 点击菜单栏 File-Read Text Data → 在 Files of Type 下拉列表中选择 xls 格式 → 选择之前保存的“1.1 均值方差分析.xls” 文档 → 在弹出的对话框中默认其选择,只需点击“Continue”。

导入成功后,将数据窗口中的文档保存为“1.1 均值方差分析.sav”,将输出窗口中的文档保存为“1.1 均值方差分析.spv”。

第三步改变变量类型:由于前面在数据预处理时,用简单的数字 1、2、3 来表示不同公司所在的行业,现在可以在 SPSS 中改变它们的变量类型,具体做法为:切换到“Variable View”,将变量公司的变量格式“Measure” 从“Scales” 变成“Nomial”,并且在“Values” 栏中逐个输入:





1 = 电力煤气及水的生产和供应业；

2 = 房地产业；

3 = 信息技术业。

(2) 正态性检验

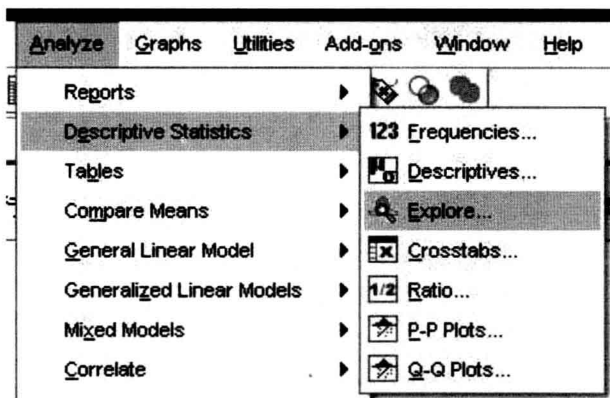
在实际工作中，人们往往很难直接判断多元数据是否来自正态总体，一种简单的办法是借助于考察每一个变量的结果来对向量的分布做出判断；并且，当数据量较大，且没有明显的证据表明所得数据不遵从多元正态时，通常认为数据来自多元正态分布总体。SPSS 软件提供了对单变量进行正态性检验的功能。

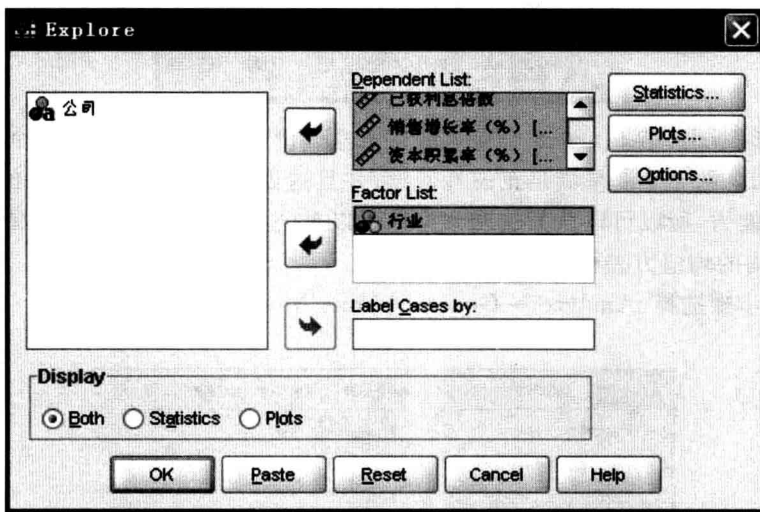
按照如下步骤选择：Analyze → Descriptive Statistics → Expolre

将 8 个指标性的变量选入 Dependent List：

在上图对话框中 Plots 选项中选择"Normality plots with tests"，然后点击 Continue，再点击 OK：

通过上述的正态性分析得到一系列的结果，其中的"Tests of Normality"是我们这一步所需要的，其他的还有很多图，他们都是具体地判断各个变量的正态性的：





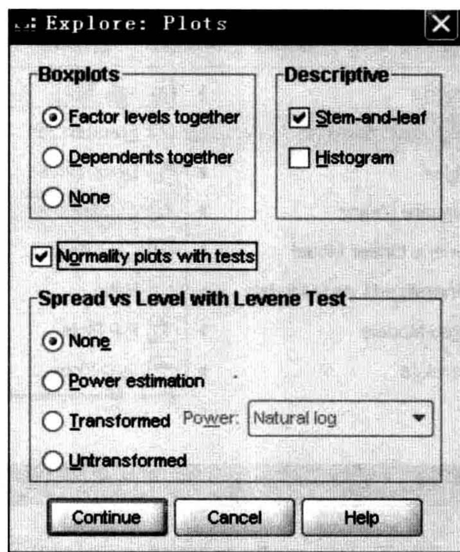
Tests of Normality

	Kolmogorov-Smirnova			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
净资产收益率(%)	0.152	35	0.039	0.944	35	0.077
总资产报酬率(%)	0.137	35	0.095	0.942	35	0.064
资产负债率(%)	0.144	35	0.065	0.939	35	0.052
总资产周转率	0.235	35	0.000	0.683	35	0.000
流动资产周转率(%)	0.159	35	0.026	0.850	35	0.000
已获利息倍数	0.172	35	0.011	0.880	35	0.001
销售增长率(%)	0.116	35	0.200 *	0.982	35	0.836
资本积累率(%)	0.252	35	0.000	0.695	35	0.000

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

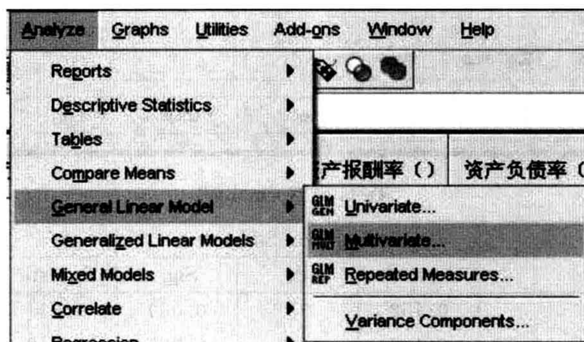
此时，由于我们的样本容量只有35个，远远小于Kolmogorov-Smirnov检验所需要的2000个样本容量的要求，所以只需要看Shapiro-Wilk统计量的值和它的Sig.。



从表中可以看出，第4、5、6、8个变量的正态性假设被拒绝了，所以这4个变量不满足正态性假设。而剩余的4个变量满足正态性假设，并且这四个指标涉及了公司的获利能力、资本结构及成长能力，所以可以认为这四个指标可以对公司运营能力作出近似的度量。

(3) 行业间的均值方差检验

按照如下步骤选择：Analyze → General Linear Model → Multivariate：



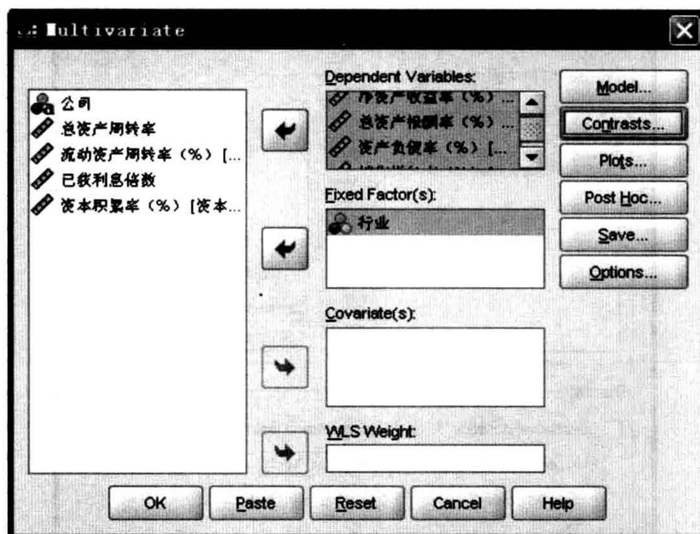
将上面确定的4个变量选入 Dependent Variables，将“行业”选入 Fixed Fator：

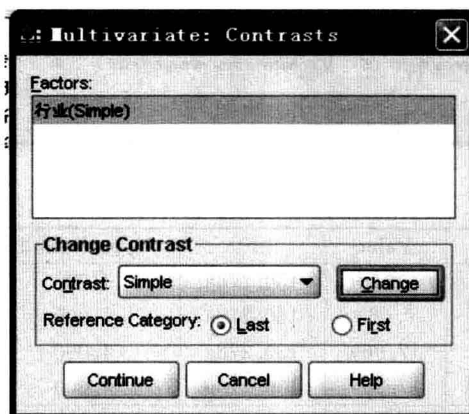
Dependent Variables 表示将要分析的变量，而将“行业”选入 Fixed Fator 是为了后面对这些指标变量在不同行业的差异进行比较分析。

点击 Contrasts 选项，在 Contrast 下拉列表中选择“Simple”，并且点击 Change，然后点击 Continue：

Contrasts 选项中的这些分析用于具体比较检验每个行业与行业之间各个指标变量的差别。

点击 Options 选项，将 Estimated Marginal Means 中的“行业”变量移到右边，选中“Compare main effects”复选框，在 Display 下面选中“Homogeneity tests”，点击 Continue，





然后点击 OK：

Options 选项中这两个选择分别用来估计各个行业内每个指标的均值和方差，以及检验它们之间的差别是否显著。

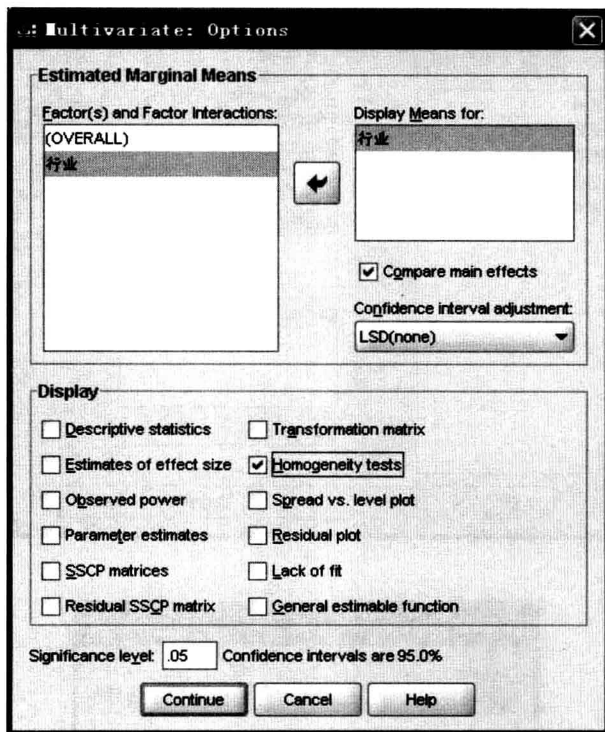
（三）主要结果及分析

（1）结果 1-1 ～ 3：总体分析行业间的 4 个指标是否存在差异，即向量的均值检验。

结果 1-1　Between-Subjects Factors

		Value Label	N
行业	1	电力煤气及水的生产和供应业	11
	2	房地产业	15
	3	信息技术业	9

上表是样本数据分别来自三个行业的个数。



结果 1-2　Multivariate Testsc

Effect		Value	F	Hypothesis df	Error df	Sig.
	Pillai's Trace	.947	130.278a	4.000	29.000	0.000
	Wilks' Lambda	0.053	130.278a	4.000	29.000	0.000
	Hotelling's Trace	17.969	130.278a	4.000	29.000	0.000
	Roy's Largest Root	17.969	130.278a	4.000	29.000	0.000
行业	Pillai's Trace	0.712	4.149	8.000	60.000	0.001
	Wilks' Lambda	0.388	4.387a	8.000	58.000	0.000
	Hotelling's Trace	1.317	4.611	8.000	56.000	0.000
	Roy's Largest Root	1.077	8.079b	4.000	30.000	0.000

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c. Design：Intercept ＋ 行业

该表给出了四个种多元检验方法，一般它们的结果都是相同的，如果不同，一般以 Hotelling's Trace 方法的结果为准。

由 Sig. 值可以看到，无论从哪个统计量来看，三个行业的运营能力（从净资产收益率、总资产报酬率、资产负债率及销售增长率这四个指标的整体来看）是有显著差别的。

实际上，GLM 模型是拟合了下面的模型：

$$\boldsymbol{Y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X + \boldsymbol{\varepsilon}$$

其中，Y ＝（净资产收益率总资产报酬率 资产负债率 销售增长率）

X = 行业

上面 Multivariate Tests 表实际上就是对该线性模型显著性的检验，此处有常数项，是因为不能肯定模型过原点。而模型通过了显著性检验，也就意味着行业的不同取值对 Y 的取值有显著影响，也就是说不同行业的运营能力是不同的。

结果 1-3 Tests of Between-Subjects Effects

Source	Dependent Variable	Type Ⅲ Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	净资产收益率(%)	458.258[a]	2	229.129	6.841	0.003
	总资产报酬率(%)	250.101[b]	2	125.050	10.034	0.000
	资产负债率(%)	1728.665[c]	2	864.333	4.515	0.019
	销售增长率(%)	9467.268[d]	2	4733.634	3.814	0.033
Intercept	净资产收益率(%)	3633.329	1	3633.329	108.483	0.000
	总资产报酬率(%)	1987.132	1	1987.132	159.453	0.000
	资产负债率(%)	71640.788	1	71640.788	374.189	0.000
	销售增长率(%)	15289.807	1	15289.807	12.321	0.001
行业	净资产收益率(%)	458.258	2	229.129	6.841	0.003
	总资产报酬率(%)	250.101	2	125.050	10.034	0.000
	资产负债率(%)	1728.665	2	864.333	4.515	0.019
	销售增长率(%)	9467.268	2	4733.634	3.814	0.033
Error	净资产收益率(%)	1071.745	32	33.492		
	总资产报酬率(%)	398.790	32	12.462		
	资产负债率(%)	6126.596	32	191.456		
	销售增长率(%)	39711.458	32	1240.983		
Total	净资产收益率(%)	4814.448	35			
	总资产报酬率(%)	2483.797	35			
	资产负债率(%)	85553.314	35			
	销售增长率(%)	60514.046	35			
Corrected Total	净资产收益率(%)	1530.003	34			
	总资产报酬率(%)	648.891	34			
	资产负债率(%)	7855.261	34			
	销售增长率(%)	49178.726	34			

a. R Squared = .300 (Adjusted R Squared = .256)

b. R Squared = .385 (Adjusted R Squared = .347)

c. R Squared = .220 (Adjusted R Squared = .171)

d. R Squared = .193 (Adjusted R Squared = .142)

上表实际上是两个一元方差分析表的合并，即分别考虑二个应变量时的方差分析结果。上面的多元方差分析已经得知行业对应变量有影响，从现在的分析表就可以更清楚地知道是对哪些自变量影响较大。由该表可以看到，四个指标的 Sig. 值分别为 0.003，0.000，0.019 及 0.033，说明三个行业在四个财务指标上均有显著差别。

（2）结果 1-4 ～ 6：每个财务指标的分析结果，同时给出了每个财务指标的方差来源。

结果 1-4　Contrast Results (K Matrix)

行业 Simple Contrast[a]			Dependent Variable			
			净资产收益率(%)	总资产报酬率(%)	资产负债率(%)	销售增长率(%)
Level 1 vs. Level 3	Contrast Estimate		−1.070	1.317	−9.215	−27.850
	Hypothesized Value		0	0	0	0
	Difference (Estimate-Hypothesized)		−1.070	1.317	−9.215	−27.850
	Std. Error		2.601	1.587	6.219	15.834
	Sig.		0.684	0.413	0.148	0.088
	95% Confidence Interval for Difference	Lower Bound	−6.368	−1.915	−21.883	−60.102
		Upper Bound	4.229	4.549	3.453	4.402
Level 2 vs. Level 3	Contrast Estimate		−7.855	−4.584	7.286	−40.942
	Hypothesized Value		0	0	0	0
	Difference (Estimate-Hypothesized)		−7.855	−4.584	7.286	−40.942
	Std. Error		2.440	1.488	5.834	14.853
	Sig.		0.003	0.004	0.221	0.010
	95% Confidence Interval for Difference	Lower Bound	−12.825	−7.616	−4.598	−71.197
		Upper Bound	−2.885	−1.552	19.170	−10.686

a. Reference category = 3

输出结果 1-4 表示，在 0.05 水平下，第一行业（电力、煤气及水的生产和供应业）与第三行业（信息技术业）各财务指标均无明显差别，说明电力、煤气及水的生产和供应业与信息技术业运营能力在统计意义上无显著差别。但由表中的第一栏可以看到，电力、煤气及水的生产和供应业的净资产收益率，资产负债率及销售增长率均低于信息技术业，总资产报酬率高于信息技术业，似乎说明信息技术业作为新生行业，其成长能力要更高一些。

第二行业（房地行业）与第三行业的净资产收益率、总资产报酬率及销售增长率三个指标有明显的差别，且在这三个指标上第三行业均大于第二行业。说明信息技术业在获利能力及成长能力上高于房地行业，而同时信息技术业的负债率较低，因此整体看来信息技术业的运营能力要高于房地行业。

结果 1-5　Multivariate Test Results

	Value	F	Hypothesis df	Error df	Sig.
Pillai's trace	0.712	4.149	8.000	60.000	0.001
Wilks' lambda	0.388	4.387a	8.000	58.000	0.000
Hotelling's trace	1.317	4.611	8.000	56.000	0.000
Roy's largest root	1.077	8.079b	4.000	30.000	0.000

该表是上面多重比较可信性的度量，由 Sig. 值可以看到，比较检验是可信的。

结果 1-6 Univariate Test Results

Source	Dependent Variable	Sum of Squares	df	Mean Square	F	Sig.
Contrast	净资产收益率(%)	458.258	2	229.129	6.841	0.003
	总资产报酬率(%)	250.101	2	125.050	10.034	0.000
	资产负债率(%)	1728.665	2	864.333	4.515	0.019
	销售增长率(%)	9467.268	2	4733.634	3.814	0.033
Error	净资产收益率(%)	1071.745	32	33.492		
	总资产报酬率(%)	398.790	32	12.462		
	资产负债率(%)	6126.596	32	191.456		
	销售增长率(%)	39711.458	32	1240.983		

该表是对每一个指标在三个行业比较的结果，与上面 Tests of Between-Subjects Effects 表中有关结果一致。

(3) 结果 1-7 ～ 10：具体估计各个变量在不同行业的均值和方差：

结果 1-7 Box's Test of Equality of Covariance Matricesa

Box's M	35.152
F	1.410
df1	20
df2	2585.573
Sig.	.106

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.
a. Design：Intercept + 行业

Levene's Test of Equality of Error Variancesa

	F	df1	df2	Sig.
净资产收益率(%)	0.585	2	32	0.563
总资产报酬率(%)	0.651	2	32	0.528
资产负债率(%)	3.448	2	32	0.044
销售增长率(%)	3.496	2	32	0.042

Tests the null hypothesis that the error variance of the dependent variable is equal across groups. a. Design：Intercept + 行业

上表第一张表是协方差阵相等的检验，检验统计量是 Box's M，由 Sig. 值可以看到，可以认为三个行业(总体)的协方差阵是相等的。第二张表给出了各行业同一指标的方差齐性检验，在 0.05 水平下，净资产收益率及总资产报酬率的方差是齐性的，而资产负债率与销售增长率的方差不齐性。这似乎说明，除了行业因素外，对资产负债率与销售增长率变动有显著影响的尚有其他因素。这与此处均值比较没有太大的关系。

结果 1-8 Estimates

Dependent Variable	行业	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
净资产收益率(%)	电力煤气及水的生产和供应业	12.320	1.745	8.766	15.874
	房地产业	5.535	1.494	2.491	8.579
	信息技术业	13.390	1.929	9.460	17.319
总资产报酬率(%)	电力煤气及水的生产和供应业	10.108	1.064	7.940	12.276
	房地产业	4.207	.911	2.351	6.064
	信息技术业	8.791	1.177	6.394	11.188
资产负债率(%)	电力煤气及水的生产和供应业	37.675	4.172	29.177	46.173
	房地产业	54.176	3.573	46.899	61.453
	信息技术业	46.890	4.612	37.495	56.285
销售增长率(%)	电力煤气及水的生产和供应业	16.445	10.622	−5.190	38.081
	房地产业	3.354	9.096	−15.173	21.881
	信息技术业	44.296	11.743	20.377	68.214

该表给出了每一行业各财务指标描述统计量的估计

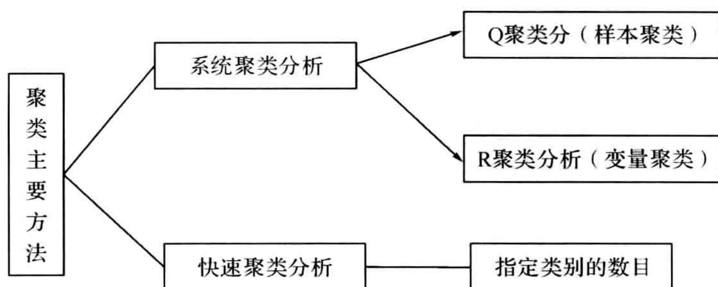
结果 1-9 Pairwise Comparisons(略)

结果 1-10 Multivariate Tests(略)

上面两个表格给出了各个指标变量在不同行业间的具体差值估计与检验。

实验二　聚类分析

2.1　实验背景



本实验利用 SPSS 自带的数据文件 World95.sav,以涉及社会,教育和经济方面的五个变量:

(1)“Urban”(城市人口比例);

(2)“Lifeexpf”(女性平均寿命)

(3)“Lifeexpm”(男性平均寿命)

(4)“Literacy”(有读写能力的人所占比例)

(5)“Gdp-cap”(人均国内总产值)

的统计数据,对 109 个国家进行分类研究,演示系统聚类分析(Q 型)和快速聚类分析的方法。

2.2　实验步骤和结果分析

(一)系统聚类法实验数据

例 2.1 为了研究世界各国的经济发展水平和文化教育水平,以便于对国家进行分类研究,这里我们进行系统聚类分析,数据为 SPSS 自带的数据文件 World95.sav。

(二)系统聚类法实验步骤

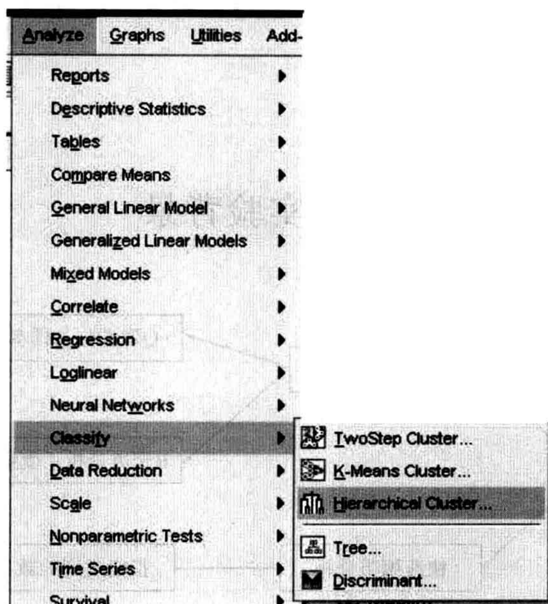
1. 打开数据

找到 SPSS 程序的安装目录,在其中的 Samples 文件夹中找到 World95.sav,双击打开,

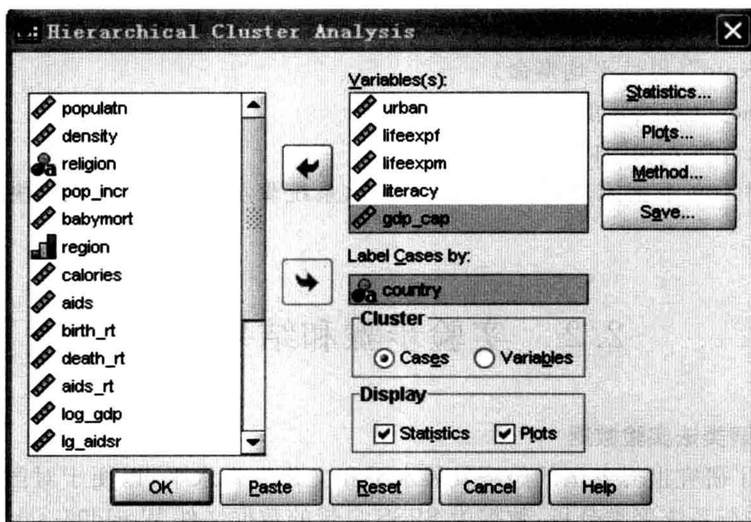
在打开成功后将数据文件另存为“系统聚类法.sav”。

2. 系统聚类法分析

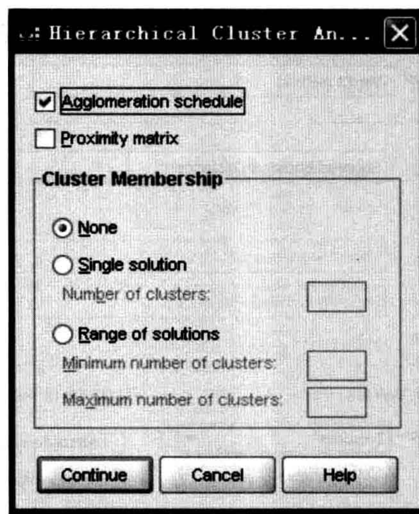
按照如下的步骤选择：Analyze → Classify → Hierarchical cluster：



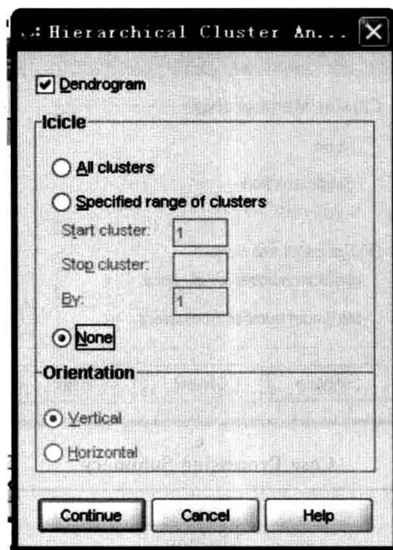
将“Urban”、“Lifeexpf”、“Lifeexpm”、“Literacy” 和“Gdp-cap” 五个变量选入 Variables 作为用于聚类分析的主要变量，将“Country” 选入 Label Cases by 中表示分类是按照不同国家来表示：



在 Statistics 选项中，默认选择“Agglomeration schedule”，它用于显示聚类的整个分析过程；“Cluster Membership” 中默认选择“None”，具体分类的个数可以在后面判断。点击 Continue：



在 Plots 选项中选择“Dendrogram”(谱系图),它可以用来分析具体分几类,当选择了类数以后每类有哪些元素。在“Icicle”(冰柱图)中选择“None”,点击 Continue:

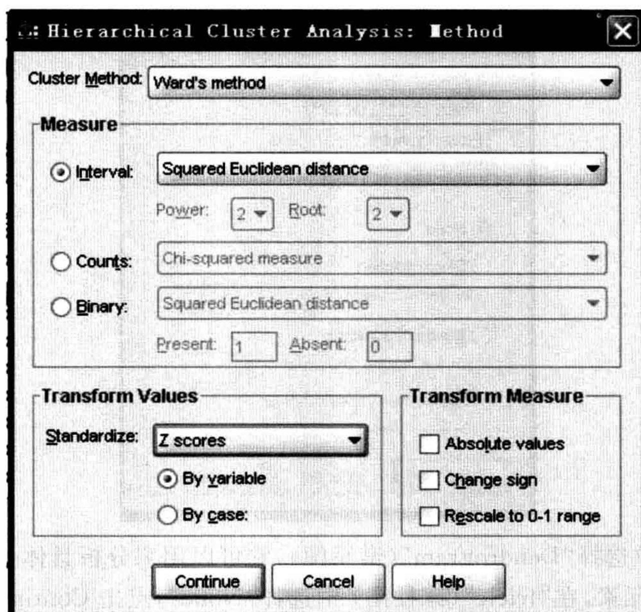


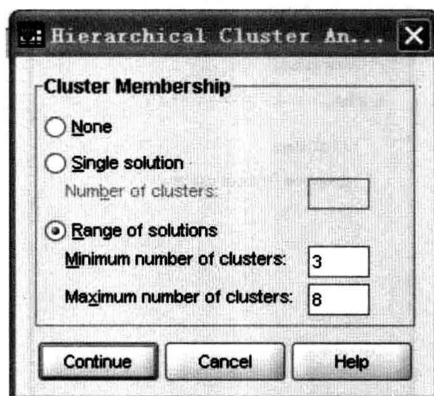
在 Method 选项中,在“Cluster Method”下拉列表中选择“Ward's method”,表示在定义类间距离时用 Ward 方法,在“Measure”下拉列表中默认“Squared Euclidean distance”,在“Transform Values”的“Standardize”的下拉列表中选择“Z-scores”将各个变量进行标准化,以消除各个变量量纲不同对距离的影响。点击 Continue:

在 Save 选项中,选择希望保存的聚类类别数范围为 3-8,点击 Continue,点击 OK:

(三)系统聚类法结果分析

(1) 结果 2-1:案例概况图:对缺失案例的分析。





Case Processing Summary[a]

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
107	98.2%	2	1.8%	109	100.0%

a. Squared Euclidean Distance used

从中发现有两个案例存在缺失数据，再看数据文档发现：第 29 位的 Czech Rep. 和第 75 位的 Oman 的 Literacy 变量没有数值，所以后面在分析的案例真正有效的只有 109－2＝107 个。

(2) 结果 2-2：聚类过程图：从中可以看出聚类的整个过程，也可以利用它来判断总共可以分多少类：

结果 2-2 Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	29	41	0.006	0	0	34
2	33	45	0.017	0	0	37
3	48	69	0.027	0	0	20
4	4	91	0.039	0	0	14
5	60	82	0.057	0	0	64
6	34	64	0.081	0	0	18
7	37	73	0.105	0	0	17
8	10	47	0.130	0	0	32
9	6	81	0.158	0	0	48
10	23	103	0.188	0	0	27
…	…	…	…	…	…	…
100	12	32	67.865	93	75	104
101	2	9	74.483	96	85	103
102	1	20	85.637	99	88	106
103	2	7	97.060	101	97	104
104	2	12	131.340	103	100	105
105	2	4	239.354	104	98	106
106	1	2	530.000	102	105	0

上述图中各项的解释：

第一列(Stage) 表示聚类分析的步数；

第二列，第三列(Cluster Combined) 表示这一步聚类中哪两个国家或小类聚成一类；

第四列(Coefficients) 表示这一步中合并的两类之间的距离；

第五列和第六列(Stage Cluster First Appear) 表示这一步聚类中参与聚类的是国家还是小类，0 表示国家，非 0 表示由第 k 步聚类生成的小类参与本步聚类；

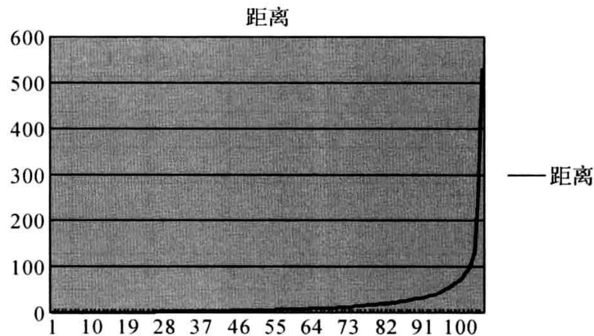
第七列(Next Stage) 表示本步聚类的结果将在以下第几步中用到。

例如：第一步中合并的分别是第 29 和 41 个案例；他们之间的距离就是 0.006；第一步中由于 29 和 41 都没有合并过，所以它们的前一步合并都是 0；第一步中合并完成的新类在第 34 步中又发生了合并，所以这里显示 34。

聚类系数 Coefficients 的作用：聚类系数表示这一步当中合并的两类的距离，并且根据系统聚类法的方法原理知道，这个距离是这一步合并前存在的所有类中距离最小的，所以聚类系数关于步数有一个递增的趋势，下图是聚类系数"：

从图中我们可以看出，开始时聚类系数递增的增量比较缓慢，说明此时合并的类别之间的距离非常小，它们理应合并为同一类；但是如果某一步合并的时候聚类系数突然太大，说明这一步中将距离相对很大的两类都归为了一类，那么这样做是不合理的，所以合理的分类应该是在前一步停止。

例如第 104 步中，聚类系数突然增大了超过 34，而之前最大的增速也只是 11 左右，所以应该在第 103 步合并完以后就停止。由于总共有 107 个有效案例，合并次数是 103，所以总共的分类为 107 − 103 = 4。



当然，关于系数突然增加多少时可以看成是一个很大的变换，这是一个主观的问题，例如上面还可以说第 103 的超过 11 的增量是突变，所以应该分为 5 类；也可以说第 105 的超过 100 的增量是突变，所以应该分为 3 类。

(3) 结果 2-3：每个案例在选择不同分类数(3-8) 时的具体归属：

结果 2-3 Cluster Membership

Case	8 Clusters	7 Clusters	6 Clusters	5 Clusters	4 Clusters	3 Clusters
1:Afghanistan	1	1	1	1	1	1
2:Argentina	2	2	2	2	2	2
3:Armenia	2	2	2	2	2	2
4:Australia	3	3	3	3	3	3
5:Austria	3	3	3	3	3	3
6:Azerbaijan	2	2	2	2	2	2
…	…	…	…	…	…	…
105:Venezuela	2	2	2	2	2	2
106:Vietnam	6	6	5	5	4	2
107:Zambia	1	1	1	1	1	1

上表中第二列给出了当总共分为 8 类时，每个有效样本他们所在的分类。后面几类可以同样解释。

(4) 结果 2-4：谱系图：从中可以非常直观地看出分几类、每类哪些元素：

结果 2-4：

谱系图给出了聚类全过程的直观表示。

在图中，将最大的类间距离算作相对距离 25，然后把其他的距离都换算成与之相对应的距离。根据研究的目的，结合聚合系数的分析，不妨之上往下分别称为第 1、2、3、4 类：

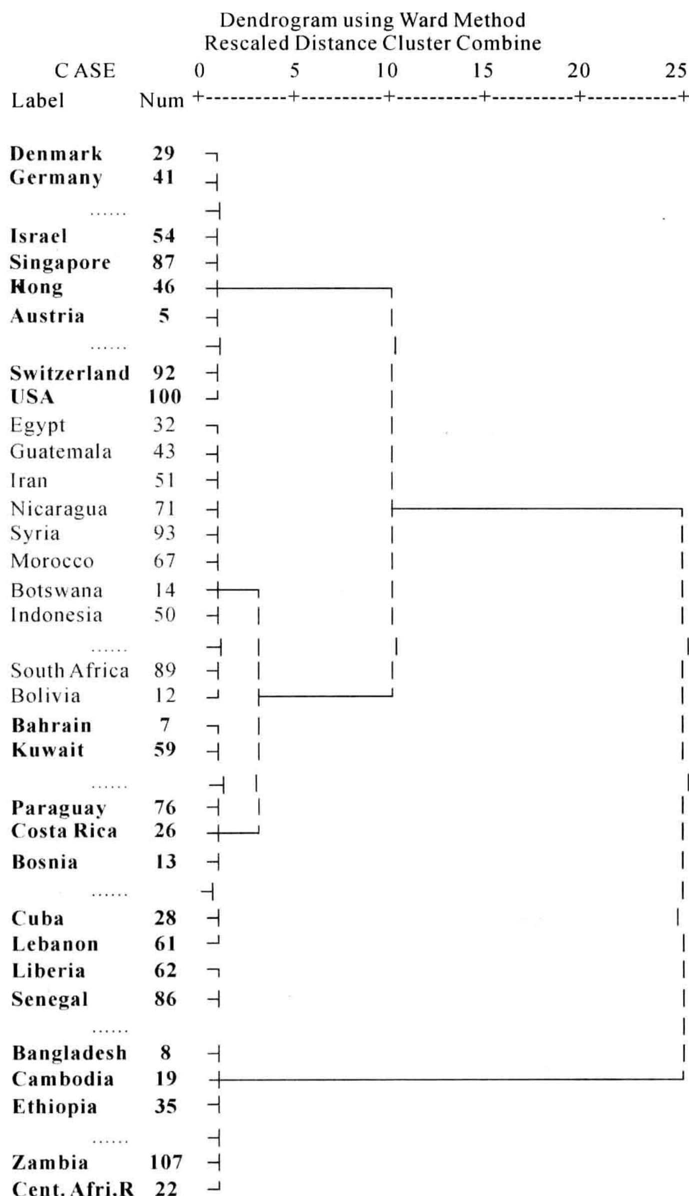
第 1 类非常明显的是发达国家和地区；

第 2 类是以中国(China) 为代表的发展中国家；

第 3 类是以俄罗斯(Russia) 为代表的发展中国家；

第 4 类是以索马里(Somalia) 为代表的最贫困的发展中国家。

当然，若只想把 107 个国家分为两类，那么可以把上面的第一类和第二类合并在一起，这是通过谱系图中连接它们的桥线看出来的，此时所有国家就分为中等发展中国家以上的国家(能够解决温饱问题) 和非常贫困的发展中国家(不能解决温饱问题)。



(四) 快速聚类法实验数据

例 2.2 为了研究亚洲国家的经济发展水平和文化教育水平，试图将亚洲国家或地区按经济和文教水平分为三类，这里我们进行快速聚类分析。数据为 SPSS 自带的数据文件 World95. sav，此时用的聚类方法为快速聚类法。

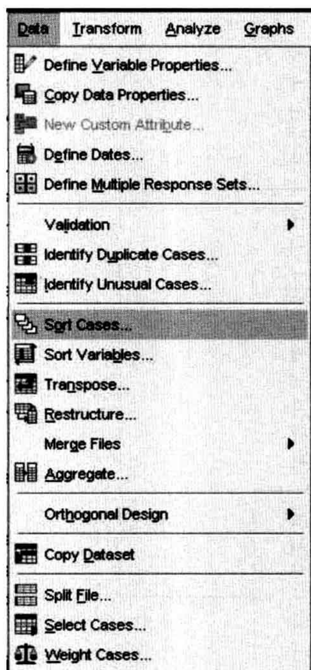
(五) 快速聚类法实验步骤

1. 打开数据

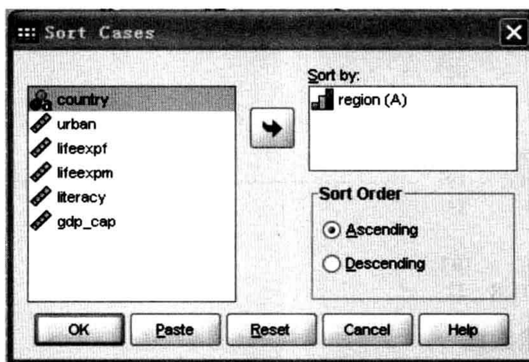
找到 SPSS 程序的安装目录，在其中的 Samples 文件夹中找到 World95. sav，双击打开，在打开成功后将数据文件另存为“快速聚类法. sav”。

2. 筛选数据

i. 首先将数据排序：Data → Sort Cases：



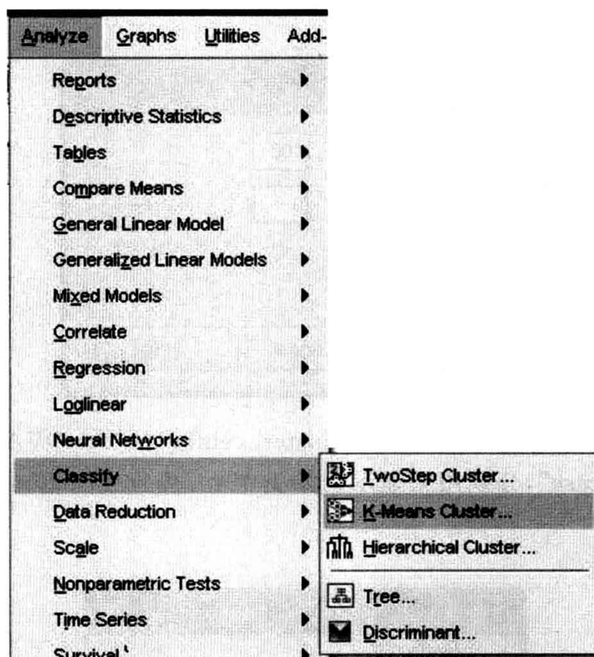
ii. 将地区变量"region"选入 Sort by，然后点击 OK：



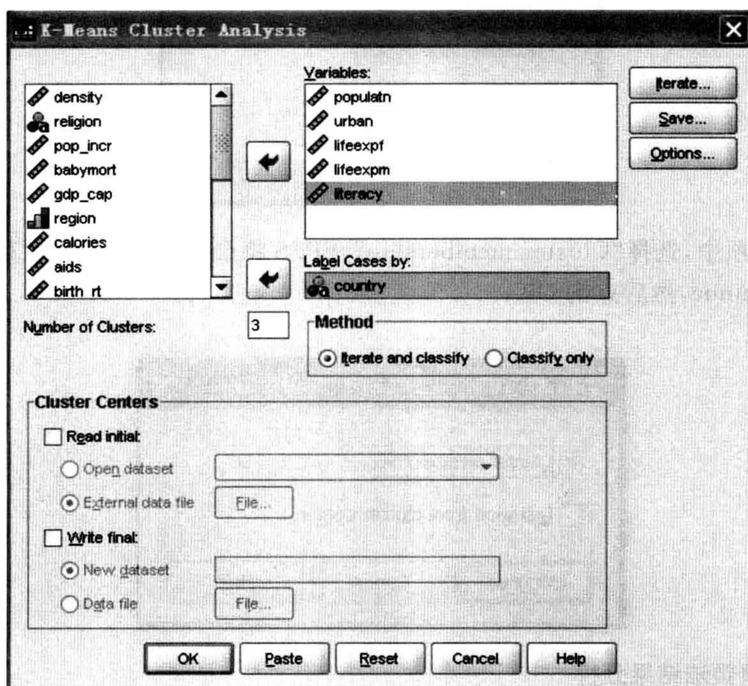
iii. 从排序完的数据窗口中将"region"取值不为 3 的所有案例全部删除，删除成功后保存文档。

3. 快速聚类法分析：

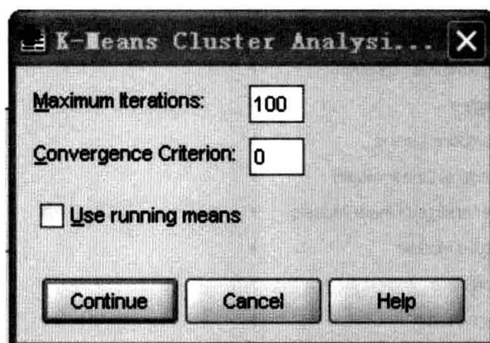
按照如下的步骤选择：Analyze → Classify → K-Means cluster：



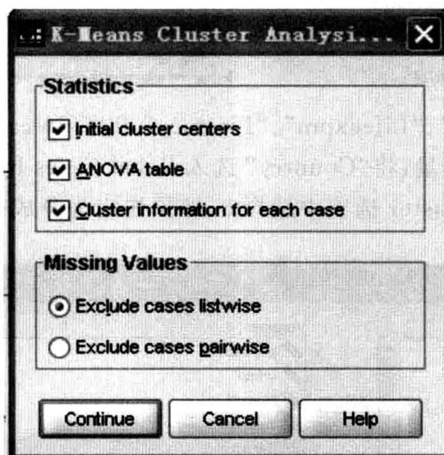
将“Urban”、“Lifeexpf”、“Lifeexpm”、“Literacy”和“Gdp-cap”五个变量选入 Variables 作为用于聚类分析的主要变量，将“Country”选入 Label Cases by 中表示分类是按照不同国家来表示，在 Number of Cluster 选项中选择 3，即预先定为分成 3 类：



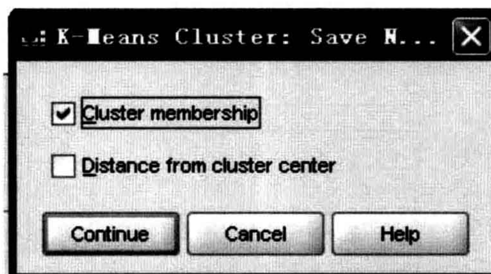
在 Iterate 选项中，选择“Maximum Iterations”为 100 次，点击 Continue：



在 Options 选项中，选择“Initial cluster center”、“ANOVA table”和“Cluster information for each case”，分别用来显示初始类中心、检验类间差异和具体分类信息。点击 Continue：



在 Save 选项中，选择“Cluster membership”，SPSS 将会保存每个样本归为哪一类这个信息。点击 Continue，然后点击 OK：



(六) 快速聚类法结果分析

(1) 结果 2-5：最初各类的重心：

结果 2-5 **Initial Cluster Centers**

	Cluster		
	1	2	3
People living in cities (%)	18	77	71
Average female life expectancy	44	82	78
Average male life expectancy	45	76	72
People who read (%)	29	99	91
Gross domestic product / capita	205	19860	7055

上表的结果非常直观,就不做解释了。

(2) 结果 2-6:样品分类情况:

结果 2-6 **Cluster Membership**

Case Number	country	Cluster	Distance
1	Afghanistan	1	571.615
2	Bangladesh	1	573.924
3	Cambodia	1	516.229
4	China	1	398.151
5	Hong Kong	2	1856.036
6	India	1	500.047
7	Indonesia	1	94.543
8	Japan	2	3363.045
9	Malaysia	1	2220.274
10	N. Korea	1	230.069
11	Pakistan	1	370.165
12	Philippines	1	96.542
13	S. Korea	3	214.034
14	Singapore	2	1507.033
15	Taiwan	3	214.034
16	Thailand	1	1025.608
17	Vietnam	1	545.396

从"Cluster Membership"全表中可以看到每个国家具体的分类。其中 Afghanistan、Bangladesh、Cambodia、China、India、Indonesia、Malaysia、N. Korea、Pakistan、Philippines、Thailand 和 Vietnam 这些发展中国家被分为了第一类。这些国家或地区的经济水平和文教水平都相对较低。这个结论可以结合下面的结果 2-7 来解释,从中可以看出第一类的各项指标都非常明显地低于其他两类。

第二分类中的国家或地区包括 Hong Kong、Japan 和 Singapore,从结果 2-7 可以看出,这三个国家或地区的所有指标都是最高的,它们是亚洲经济和文教水平最发达的地方。

第三分类包括 S. Korea 和 Taiwan,它们是介于中间的国家和地区,比发展中国家整体水平高,但是比不上 Hong Kong、Japan 和 Singapore。

(3) 结果 2-7:最后各类的重心:

结果 2-7　Final Cluster Centers

	Cluster		
	1	2	3
People living in cities (%)	29	90	72
Average female life expectancy	63	80	76
Average male life expectancy	60	75	70
People who read (%)	66	88	94
Gross domestic product / capita	775	16497	6841

(4) 结果 2-8:方差分析表(类间差别检验):

结果 2-8　ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	Df		
People living in cities (%)	5336.488	2	169.577	14	31.469	0.000
Average female life expectancy	454.600	2	70.494	14	6.449	0.010
Average male life expectancy	321.326	2	41.113	14	7.816	0.005
People who read (%)	1073.096	2	570.625	14	1.881	0.189
Gross domestic product / capita	3.042E8	2	1780295.690	14	170.846	0.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

从上面的方差分析表的显著水平 Sig. 可以看出,所有 5 个变量中除了第四个变量外,其他四个变量都存在显著的差别。

实验三　判别分析

3.1　实验背景

SPSS的Discriminant模块提供三种判别分析的方法:(1)使用SPSS默认选项进行判别分析,给出的是标准化的Fisher判别函数结果;(2)使用选择项进行判别分析,给出Bayes判别的结果,但需要注意的是,输出Bayes判别的复选框的名字叫Fisher,这是因为这种思想是Fisher提出的,故而SPSS如此命名;(3)进行逐步判别分析。

本实验通过例3.1说明(1)(2)在SPSS中的实现,利用例3.2展示如何进行逐步判别分析。

3.2　实验步骤和结果分析

(一)Fisher判别法和Bayes判别法实验数据

例3.1 一个城市的居民家庭,按其有无割草机可分为两组,有割草机的一组记为π_1,没有割草机的一组记为π_2,割草机工厂欲判断一些家庭是否将购买割草机。从π_1和π_2分别随机抽取12个样品,调查两项指标:x_1 = 家庭收入;x_2房前屋后土地面积。数据如下:

有割草机家庭		无割草机家庭	
x1	x2	x1	x2
20.0	9.2	25.0	9.8
28.5	8.4	17.6	10.4
21.6	10.8	21.6	8.6
20.5	10.4	14.4	10.2
29.0	11.8	28.0	8.8
36.7	9.6	16.4	8.8
36.0	8.8	19.8	8.0
27.6	11.2	22.0	9.2
23.0	10.0	15.8	8.2
31.0	10.4	11.0	9.4
17.0	11.0	17.0	7.0
27.0	10.0	21.0	7.4

(二)Fisher 判别法实验步骤

1. 数据预处理

第一步 Excel 处理:为了便于进行 SPSS 分析,将上述原始数据的 Excel 文档改为如下形式:

y(有无割草机家庭)	x1(家庭收入)	x2(房前屋后土地面积)
1	20.0	9.2
1	28.5	8.4
1	21.6	10.8
1	20.5	10.4
1	29.0	11.8
1	36.7	9.6
1	36.0	8.8
1	27.6	11.2
1	23.0	10.0
1	31.0	10.4
1	17.0	11.0
1	27.0	10.0
0	25.0	9.8
0	17.6	10.4
0	21.6	8.6
0	14.4	10.2
0	28.0	8.8
0	16.4	8.8
0	19.8	8.0
0	22.0	9.2
0	15.8	8.2
0	11.0	9.4
0	17.0	7.0
0	21.0	7.4

即分别用数字 1、0 来代替一个家庭是否有割草机,并且将两类家庭的 x_1 和 x_2 数据合并,并将数据保存为“判别分析. xls”。

第二步数据导入:将 xls 格式的 Excel 数据导入到 SPSS 中:打开 SPSS → 点击菜单栏 File-Read Text Data → 在 Files of Type 下拉列表中选择 xls 格式 → 选择之前保存的“判别分析. xls” 文档 → 在弹出的对话框中默认其选择,只需点击“Continue”;具体步骤和前面均值方差分析一样,所以就不用图示了。导入成功后,将数据窗口中的文档保存为“判别分析. sav”。

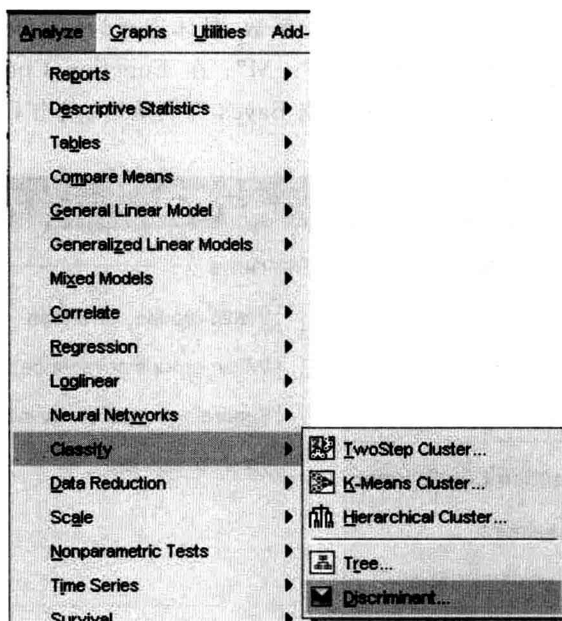
第三步改变变量类型:由于前面在数据预处理时,用简单的数字 1、0 来代替一个家庭是否有割草机,现在可以在 SPSS 中改变它们的变量类型,具体做法为:切换到“Variable View”,将变量 y 的变量格式“Measure” 从“Scales” 变成“Nomial”,并且在“Values” 栏中逐个输入:

1 = 有割草机家庭;

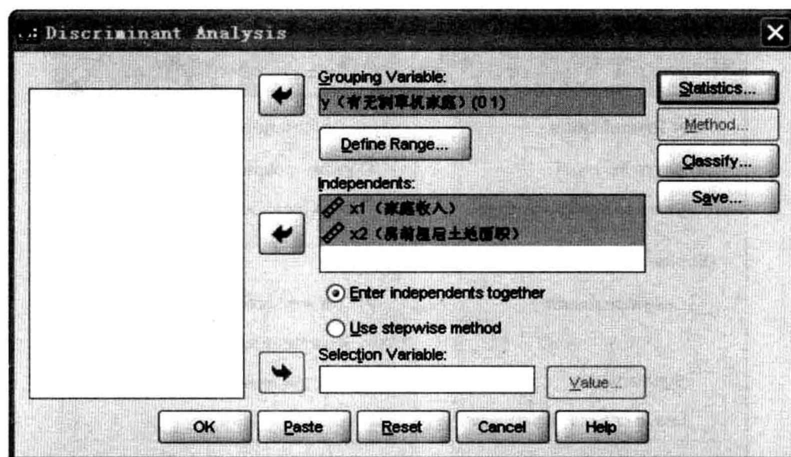
0 = 无割草机家庭。

2. Fisher 判别法与 Bayes 判别法

按照如下步骤选择：Analyze → Classify → Discriminant：



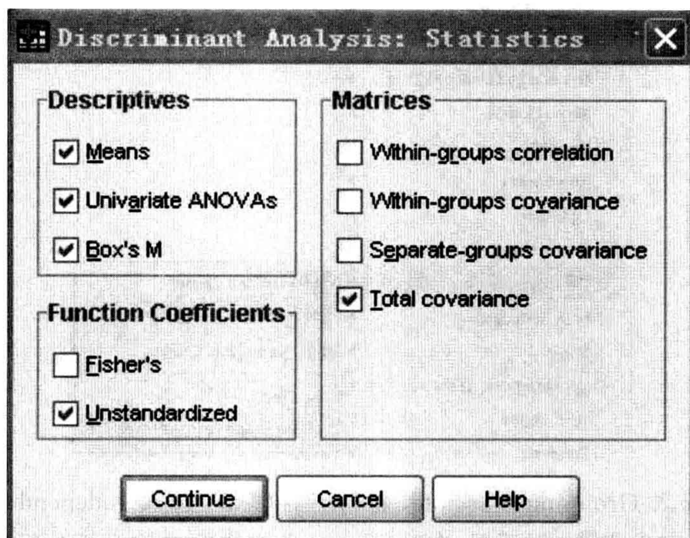
将变量“y”选入 Grouping Variable，将“x1”和“x2”选入 Independents，点击 Define Range”(判别分析定义范围）对话框，在“Minimum”文本框中输入该分组变量的最小值 0，在“Maximum”文本框中输入该分组变量的最大值 1，单击“Continue”按钮，返回主对话框



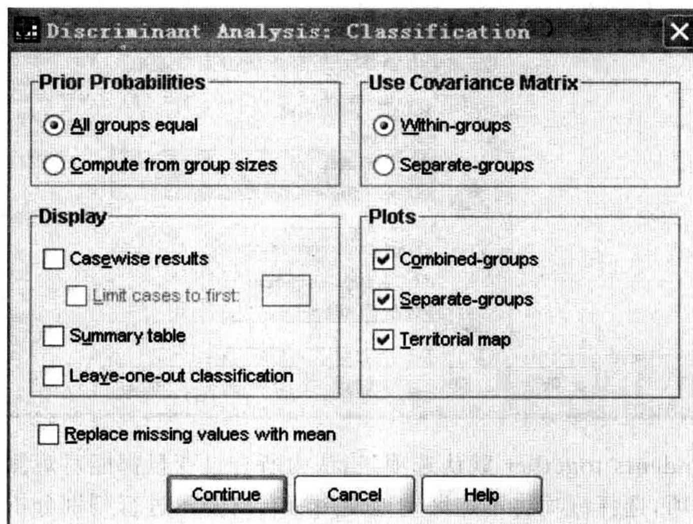
Enter independents together 默认选项。当认为所有自变量都能对观测特性提供丰富的信息时，使用该选项，选择该项将不加区别地使用所有自变量进行判别分析，建立全模型，且不需要进一步选择，这里选择默认项。

Use stepwise method 逐步分析方法。当认为不是所有自变量都能对观测量特性提供丰富的信息时，选择该项，因此需要判别贡献的大小再进行选择。选中该单选按钮时，“Method”按钮被激活，可以进一步选择判别分析方法。

在 Statistics 选项中，可以选择很多对于 Fisher 判别法适用条件的检验，例如方差齐性等。同时，还可以给出 Fisher 判别的非标准化系数。具体选择为：在 Descriptives 选项下选择“Means”、“Univariate ANOVAs”和“Box's M”；在 Function Coefficients 选项下选择 Unstanardized。注意，选项 Fisher's 表示输出 Bayes 判别函数。点击 Continue：



在 Classify 选项中，可以选择将判别结果进行图形化的展示，选择 Plots 下面的“Combined-groups”、“Separate-groups”和“Territorial map”。



(三)Fisher 判别法结果分析

(1) 结果 3-1:有效样本分析、均值方差适用条件检验:

结果 3-1-1　Analysis Case Processing Summary

Unweighted Cases		N	Percent
Valid		24	100.0
Excluded	Missing or out-of-range group codes	0	0.0
	At least one missing discriminating variable	0	0.0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	0.0
	Total	0	0.0
Total		24	100.0

结果 3-1-2　Group Statistics

y(有无割草机家庭)		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
无割草机家庭	x1(家庭收入)	19.133	4.7224	12	12.000
	x2(房前屋后土地面积)	8.817	1.0564	12	12.000
有割草机家庭	x1(家庭收入)	26.492	6.2596	12	12.000
	x2(房前屋后土地面积)	10.133	1.0103	12	12.000
Total	x1(家庭收入)	22.812	6.5977	24	24.000
	x2(房前屋后土地面积)	9.475	1.2141	24	24.000

结果 3-1-3　Tests of Equality of Group Means

	Wilks'Lambda	F	df1	df2	Sig.
x1(家庭收入)	.676	10.568	1	22	.004
x2(房前屋后土地面积)	.693	9.736	1	22	.005

结果 3-1-4　Log Determinants

y(有无割草机家庭)	Rank	Log Determinant
无割草机家庭	2	3.207
有割草机家庭	2	3.587
Pooled within-groups	2	3.447

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

结果 3-1-5　Test Results

Box's M		1.102
F	Approx.	0.331
	df1	3
	df2	87120.000
	Sig.	0.803

Tests null hypothesis of equal population covariance matrices.

结果 3-1-1 给出有效样本数目,从数据中可以看出所有 24 个样本都是有效的。结果 3-1-2 给出两组数据分别的均值方差估计。结果 3-1-3 给出两组间均值的比较,从显著性水平 Sig.

可以看出，两组数据的 x1 和 x2 变量的均值都存在显著的差异，所以满足判别分析所要求的适用条件。结果 3-1-4 和 3-1-5 给出了两组间变量 x1 和 x2 方差齐性的检验。从结果 3-1-5 的显著性水平可以看出，齐方差性没有被拒绝，因此符合判别分析的适用条件。

（2）结果 3-2：判别结果分析

结果 3-2-1　Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	1.167a	100.0	100.0	0.734

a. First 1 canonical discriminant functions were used in the analysis.

结果 3-2-2　Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	0.461	16.243	2	0.000

结果 3-2-3　Standardized Canonical Discriminant Function Coefficients

	Function
	1
x1（家庭收入）	0.806
x2（房前屋后土地面积）	0.785

结果 3-2-4　Canonical Discriminant Function Coefficients

	Function
	1
x1（家庭收入）	0.145
x2（房前屋后土地面积）	0.759
(Constant)	−10.508

Unstandardized coefficients

结果 3-2-5　Structure Matrix

	Function
	1
x1（家庭收入）	0.641
x2（房前屋后土地面积）	0.616

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions

Variables ordered by absolute size of correlation within function.

结果 3-2-6　Functions at Group Centroids

y（有无割草机家庭）	Function
	1
无割草机家庭	−1.034
有割草机家庭	1.034

Unstandardized canonical discriminant functions evaluated at group means

由于此时的分类变量 y 只分为 2 组，而且用于分析的自变量也只有 2 个，所以 Fisher 判别函数的个数为 $\min\{2,2-1\}=1$，即只有一个判别函数。

结果 3-2-1 给出了 Fisher 判别函数对应的特征值，由于在判别分析中，一个判别函数所代表的方差量用所对应的特征值(eigenvalue) 来相对表示，即组间偏差平方和与组内偏差平方和之比，方差越大说明分组差异越显著，即该判别函数对总的判别结果影响越明显(判别能力越强)。典型相关系数 Canonical correlations

$$\text{Can. Corr} = \sqrt{\frac{\text{Eigenvalue}_i}{1 + \text{Eigenvalue}_i}}$$

典型相关系数值越大，在这一判别轴上分组差异越明显。

结果 3-2-2 给出了 Wilks' Lambda 值，间接地进行判别函数的显著性检验，其值越小表示越高的判别力：

$$\text{Wilks' Lambda} = \prod_{i}^{I} \frac{1}{1 + \text{Eigenvalue}_i}$$

从 Sig. 可以看出第一个判别函数的特征值是显著非零的，即它的信息量是显著非零的。

结果 3-2-3 和结果 3-2-4 分别给出了标准化与未对 x1 和 x2 进行标准化的线性判别函数的判别系数。

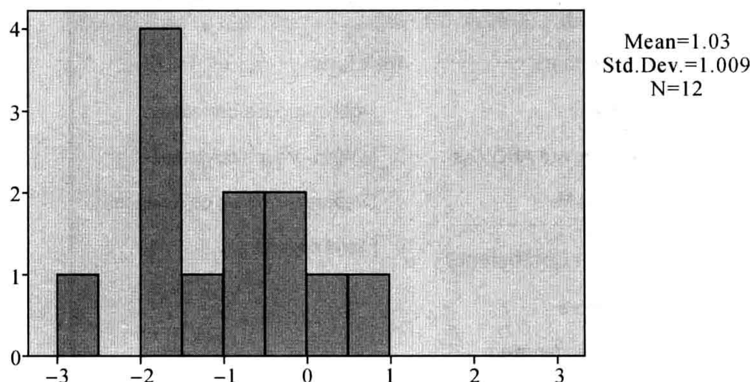
标准化判别函数为：$ZFunc = 0.806zx_1 + 0.785zx_2$

Fisher 判别函数为：$Func = 0.145x_1 + 0.759x_2 - 10.508$

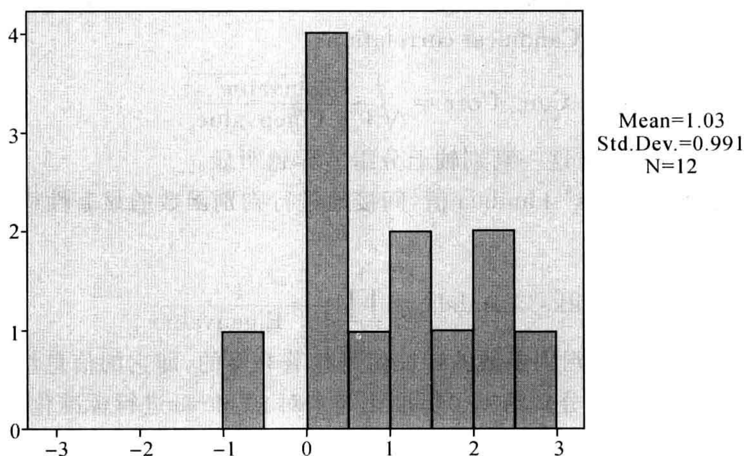
非标准化判别函数是用来计算判别值的，标准化判别系数比较各变量对判别值的相对作用程度：哪个变量的标准化系数的绝对值大，就意味着它对判别值有较大影响。从标准化判别函数可以看出，x1 因素的分组能力较 x2 因素强，若 x1 = 20，x2 = 8，那么判别的得分为：$-0.018 = 0.145 * 20 + 0.759 * 10 - 10.508$。

结果 3-2-5 给出了判别得分(即判别函数的应变量)与自变量的相关系数，它可以用来表示判别函数中各个自变量对判别得分起作用的大小。从上面可以看出，"家庭收入"对判别得分显得略微重要。

结果3-3-1:Canonicaal Discriminant Function1
y(有无割草机家庭)=无割草机家庭



结果 3-2-6 给出了由每个判别函数给出的各组的判别重心。利用 Fisher 判别函数计算出各观测值具体坐标后，再计算出离各重心的距离，则可得知分类情况，由于这里只有一个判别函数，所以对于每组只有一个判别重心。



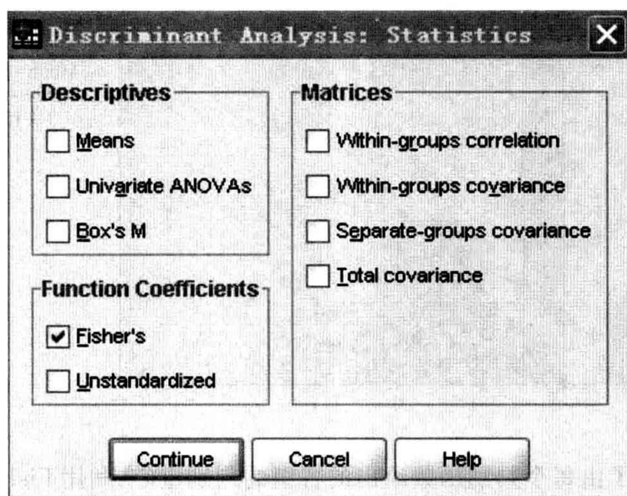
(3) 结果 3-2-7:图形化显示的判别结果:

由于此时只有一个判别函数,所以领域图(Territorial map)和联合分布图(Combined-groups)都不能显示,只能显示单独分布图(Separate-groups)的内容。当判别函数个数 = min{自变量个数,分组数-1}超过一个时,通过领域图和联合分布图可以非常直观的判断样本被判别到哪一组。

上面两个图分别给出了两组样本内样本数关于第一个判别函数值的条形图。例如结果 3-3-1 中表示判别函数取值在[−3,−2]内的样本数只有一个。

(四)Bayes 判别法实验步骤

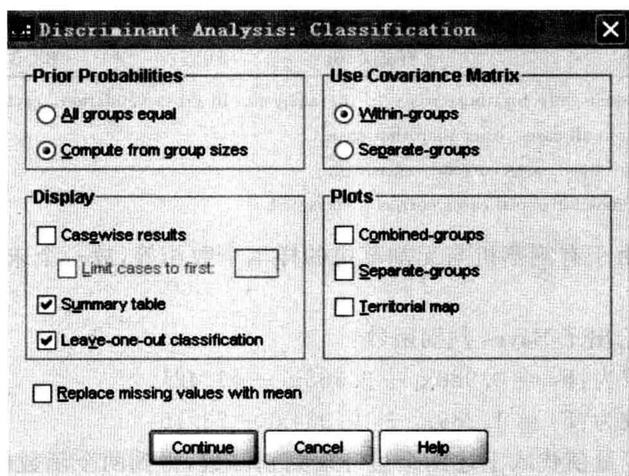
在这里,我们依然利用例 3.1 的有关数据进行实验。在 Statistics 选项中,选项 Fisher's 表示输出 Bayes 判别函数,选择该选项,点击 Continue。



在 Classify 选项中,在 Display 选项组中选择生成到输出窗中的分类结果,其中

Summary table 要求输出分类的综述表，给出正确分类观测数(原始类和根据判别函数计算的预测类相同)和错分观测量数即错分率；Leave-one-out classification 输出对每个观测量进行分类的结果，所依据的判别时由除该观测量以外的其他观测量导出的，也称为交互校验结果。点击 Continue。然后点击 OK。

在 Prior Probabilities 选项中选择先验概率，有两个单选项供选择：All groups equal 表示各类先验概率相等，Compute from groups sizes 表示由各类的样本量计算决定，即各类的先验概率与其样本量成正比，这里选择该项。



(五)Bayes 判别法结果分析

结果 3-3-1 Prior Probabilities for Groups

有无割草机	Prior	Cases Used in Analysis	
		Unweighted	Weighted
无割草机	0.500	12	12.000
有割草机	0.500	12	12.000
Total	1.000	24	24.000

结果 3-3-2 Classification Function Coefficients

		有无割草机	
		无割草机	有割草机
家庭收入	0.988	1.289	
屋前屋后的土地面积	9.363	10.934	
(Constant)	−51.421	−73.160	

Fisher's linear discriminant functions

结果 3-3-3 **Classification Resultsb,c**

		有无割草机	Predicted Group Membership		Total
			无割草机	有割草机	
Original	Count	无割草机	10	2	12
		有割草机	1	11	12
	%	无割草机	83.3	16.7	100.0
		有割草机	8.3	91.7	100.0
Cross-validated[a]	Count	无割草机	9	3	12
		有割草机	2	10	12
	%	无割草机	75.0	25.0	100.0
		有割草机	16.7	83.3	100.0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 87.5% of original grouped cases correctly classified.

c. 79.2% of cross-validated grouped cases correctly classified.

结果 3-3-1 中,由于有割草机与无割草机的样本个数相等,故一个未知样本的先验概率也相等。

结果 3-3-2 中,给出了 Bayes 判别函数:

第一类判别函数为:$F_1 = 0.988x_1 + 9.363x_2 - 51.421$

第二类判别函数为:$F_2 = 1.289x_1 + 10.934x_2 - 73.16$

将两样品的自变量值代入上述两个贝叶斯判别函数,得到两个函数值,比较这两个函数值,哪个函数值比较大就可将该样品判入该类。

结果 3-3-3 中,给出了正确、错误判别率。各组正确判别率为 83.3% 和 91.7%,交互验证法:两组的正确判别率分别为 75% 和 83.3%。

(六) 逐步判别法实验数据

例 3.2 研究者希望能够根据气候、经济因素、人口等信息来判断某国家或地区属于哪一类型。这里国家 country(因变量) 有 3 种类别,OECD 表示经合组织的国家(包括美国、加拿大和西欧等发达国家),Pacific/Asia 表示亚太地区的国家,Africa 表示非洲地区的国家。考虑了以下几个自变量,climate(气候因素,包括沙漠气候、干旱气候、地中海气候、海洋气候、温带气候和极地气候等),urban(城市居民的比例),population(人口数),gdp_cap(人均 GDP)。数据集来自 SPSS10.0 自带的数据集 World95.sav。

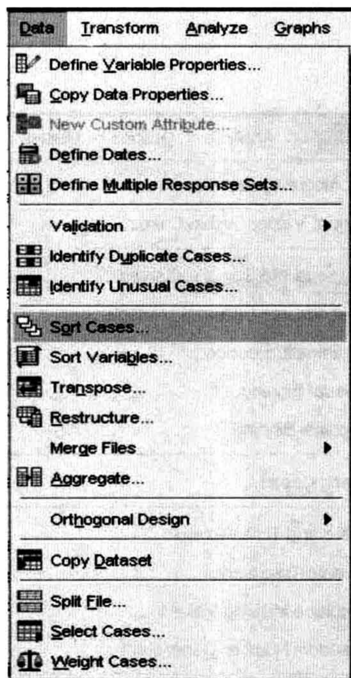
(七) 逐步判别法实验步骤

1. 打开数据

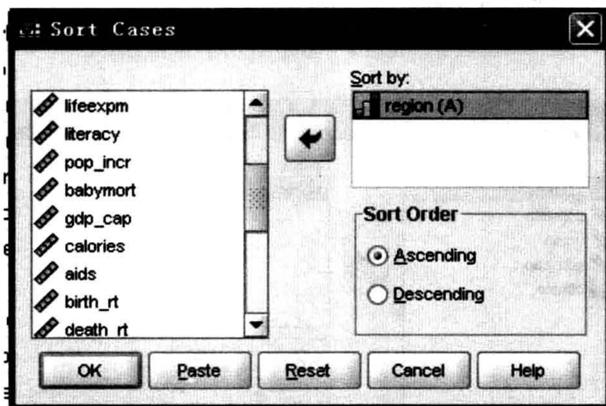
找到 SPSS 程序的安装目录,在其中的 Samples 文件夹中找到 World95.sav,双击打开,在打开成功后将数据文件另存为"逐步判别分析.sav"。

2. 筛选数据

首先将数据排序:Data → Sort Cases:

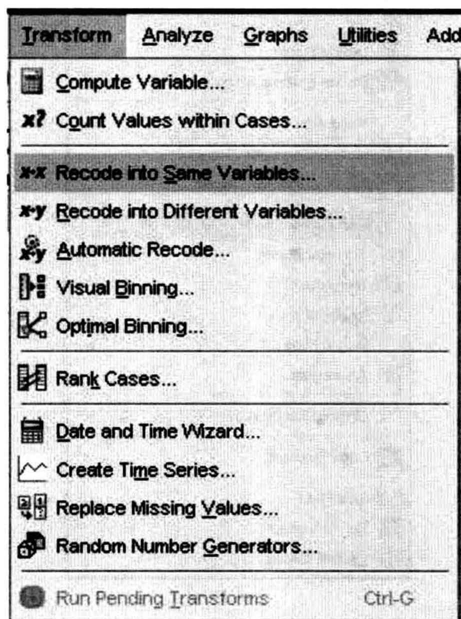


将地区变量"region"选入 Sort by，然后点击 OK：

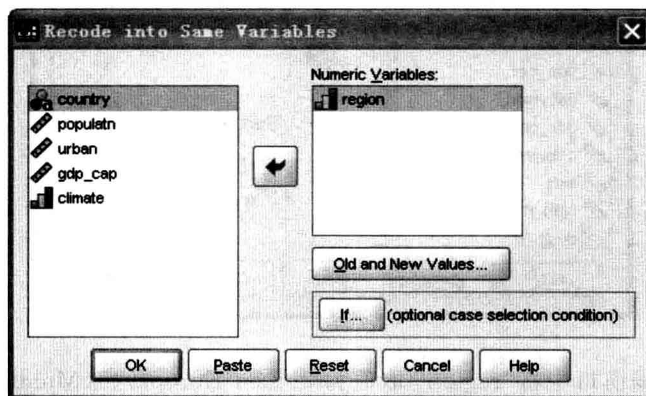


从排序完的数据窗口中将"region"取值为 2(East Europe)、5(Middle East) 和 6(Latin America) 的所有案例全部删除；然后切换到"Variable View"，将除了"country"、"population"、"urban"、"gdp_cap"、"region" 和"climate" 这些变量以外的其他所有变量删除，删除成功后保存文档；

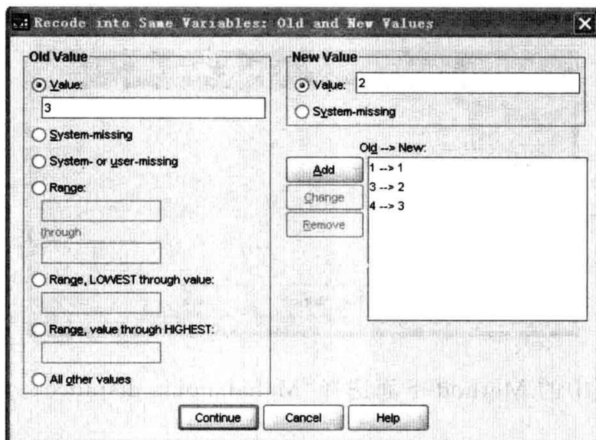
将“region”重新按 1 ～ 3 编号：首先按如下方式选择：Transform → Recode into Same Variables；



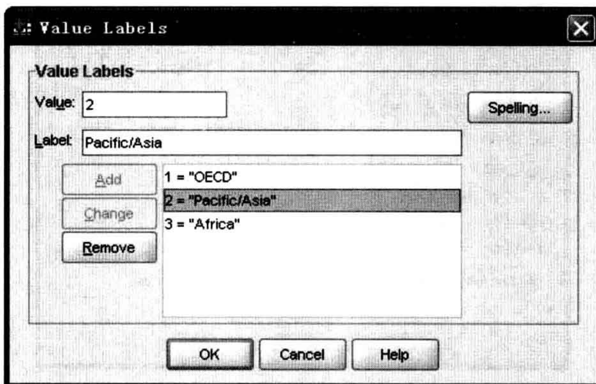
将“region”选入右边的 Numeric Variable 框内：



点击 Old and New Values 选项，在 Old Value 框内输入 3，在 New Value 框内输入 2，点击 Add，直到将所需改变的变量值都输入完成。点击 Continue。点击 OK：

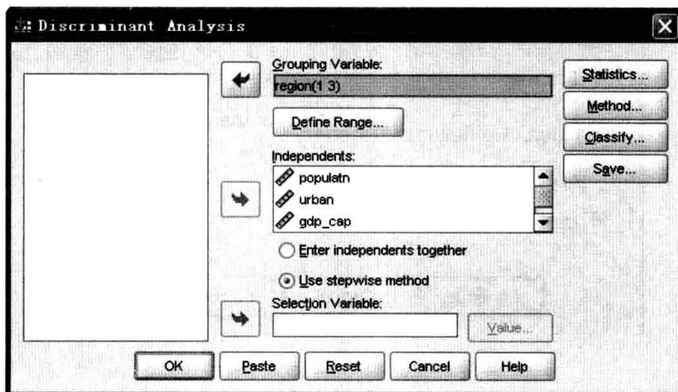


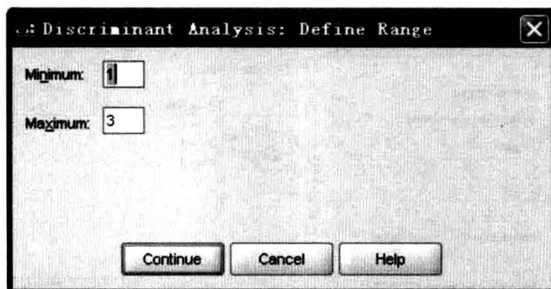
最后，在 Variable View 中将 Region 的变量值也作相应的调整：



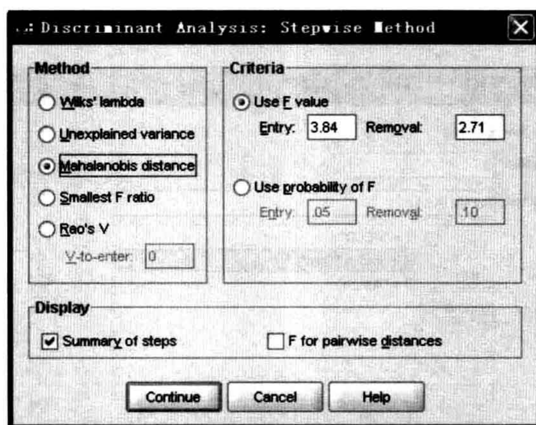
3. 逐步判别分析

按照如下方式选择：Analyze → Classify → Discriminant：将“region”变量选入 Grouping Variables，点击 Define Range，在 Minimum 内输入 1，在 Maximum 内输入 3，表示所要选择的分组从第一组到第三组；将“population”、“urban”、“gdp_cap”和“climate”这四个变量选入 Independents；选择 Use stepwise method 逐步分析方法。

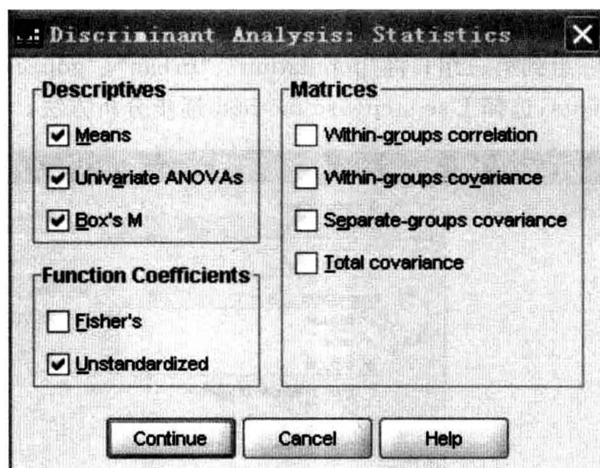


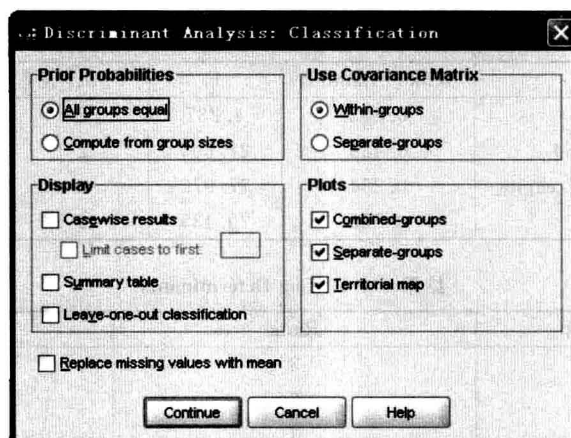


在 Methods 选项中的 Method 下面选择“Mahalanobis distance”。点击 Continue：



在 Statistics 和 Classify 选项中，选择同上面 Fisher 判别分析：





(八) 逐步判别法结果分析

(1) 结果 3-4:样本有效性分析、判别分析适用条件检验:

结果 3-4-1　Analysis Case Processing Summary

Unweighted Cases		N	Percent
Valid		56	98.2
Excluded	Missing or out-of-range group codes	0	0.0
	At least one missing discriminating variable	1	1.8
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	0.0
	Total	1	1.8
Total		57	100.0

结果 3-4-2　Group Statistics

Region or economic group		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
OECD	Population in thousands	33085.10	57148.252	21	21.000
	People living in cities (%)	74.71	14.890	21	21.000
	Gross domestic product / capita	16610.86	3725.971	21	21.000
	Predominant climate	7.76	1.261	21	21.000
Pacific/Asia	Population in thousands	189012.50	348024.915	16	16.000
	People living in cities (%)	43.12	28.654	16	16.000
	Gross domestic product / capita	4088.50	6454.734	16	16.000
	Predominant climate	5.75	1.483	16	16.000
Africa	Population in thousands	19757.11	24357.858	19	19.000
	People living in cities (%)	29.26	15.062	19	19.000
	Gross domestic product / capita	998.68	1178.258	19	19.000
	Predominant climate	4.89	1.629	19	19.000
Total	Population in thousands	73113.79	199794.360	56	56.000
	People living in cities (%)	50.27	27.825	56	56.000
	Gross domestic product / capita	7736.05	8154.118	56	56.000
	Predominant climate	6.21	1.904	56	56.000

结果 3-4-3　**Tests of Equality of Group Means**

	Wilks' Lambda	F	df1	df2	Sig.
Population in thousands	0.862	4.237	2	53	0.020
People living in cities (%)	0.489	27.665	2	53	0.000
Gross domestic product / capita	0.254	77.972	2	53	0.000
Predominant climate	0.565	20.435	2	53	0.000

结果 3-4-4　**Log Determinants**

Region or economic group	Rank	Log Determinant
OECD	3	38.594
Pacific/Asia	3	43.632
Africa	3	34.970
Pooled within-groups	3	41.667

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

结果 3-4-5　**Test Results**

Box's M		152.522
F	Approx.	11.642
	df1	12
	df2	12045.329
	Sig.	0.000

Tests null hypothesis of equal population covariance matrices.

结果 3-4-1 从中可以看出则 57 个样本国家中含有一个缺失值，这个在之前聚类分析的时候就已经确认了。

结果 3-4-2 是分组的个变量均值方差估计。

结果 3-4-4—3-4-5 给出的是齐方差性检验。

(2) 结果 3-5：逐步判别的过程和结果：自变量选取

结果 3-5-1　**Variables Entered/Removeda,b,c,d**

Step	Entered	Min. D Squared					
		Statistic	Between Groups	Exact F			
				Statistic	df1	df2	Sig.
1	Gross domestic product / capita	0.545	Pacific/Asia and Africa	4.738	1	53.000	0.034
2	Population in thousands	1.590	Pacific/Asia and Africa	6.775	2	52.000	0.002
3	Predominant climate	1.775	Pacific/Asia and Africa	4.945	3	51.000	0.004

At each step, the variable that maximizes the Mahalanobis distance between the two closest groups is entered.

a. Maximum number of steps is 8.

b. Minimum partial F to enter is 3.84.

c. Maximum partial F to remove is 2.71.

d. F level, tolerance, or VIN insufficient for further computation.

结果 3-5-2 Variables in the Analysis

Step		Tolerance	F to Remove	Min. D Squared	Between Groups
1	Gross domestic product / capita	1.000	77.972		
2	Gross domestic product / capita	0.976	76.492	0.005	OECD and Africa
	Population in thousands	0.976	4.155	0.545	Pacific/Asia and Africa
3	Gross domestic product / capita	0.976	41.680	0.978	Pacific/Asia and Africa
	Population in thousands	0.940	3.698	0.928	Pacific/Asia and Africa
	Predominant climate	0.961	5.594	1.590	Pacific/Asia and Africa

结果 3-5-3 Variables Not in the Analysis

Step		Tolerance	Min. Tolerance	F to Enter	Min. D Squared	Between Groups
0	Population in thousands	1.000	1.000	4.237	0.005	OECD and Africa
	People living in cities (%)	1.000	1.000	27.665	0.489	Pacific/Asia and Africa
	Gross domestic product / capita	1.000	1.000	77.972	0.545	Pacific/Asia and Africa
	Predominant climate	1.000	1.000	20.435	0.344	Pacific/Asia and Africa
1	Population in thousands	0.976	0.976	4.155	1.590	Pacific/Asia and Africa
	People living in cities (%)	0.545	0.545	0.791	0.620	Pacific/Asia and Africa
	Predominant climate	0.998	0.998	6.113	0.928	Pacific/Asia and Africa
2	People living in cities (%)	0.536	0.536	1.086	1.758	Pacific/Asia and Africa
	Predominant climate	0.961	0.940	5.594	1.775	Pacific/Asia and Africa
3	People living in cities (%)	0.536	0.536	1.056	1.938	Pacific/Asia and Africa

结果 3-5-4 Wilks' Lambda

Step	Number of Variables	Lambda	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
1	1	0.254	1	2	53	77.972	2	53.000	0.000
2	2	0.219	2	2	53	29.596	4	104.000	0.000
3	3	0.179	3	2	53	23.141	6	102.000	0.000

结果 3-5-1 给出了进入判别函数的 3 个自变量，以及它们进入的顺序。并且给出了它们的显著性。从上面的 Sig. 值可以看出加入这三个变量后的判别函数是显著的。

结果 3-5-2 给出了每一步中选取进入判别函数的自变量，从第一步中的 1 个自变量到第三步中的 3 个自变量。

结果 3-5-3 给出了每一步中，没有被选中的自变量。从初始第 0 步的 4 个变量到第 3 步的 1 个变量。

结果 3-5-4 对每一步得到的判别函数都进行了显著性检验。从上面的显著性水平 Sig. 可以看出，三个判别函数都是显著的。

(3) 结果 3-6：逐步判别的结果：判别函数

这里得到的结果和前面典型判别分析的结果有完全同样的解释，只不过此时有两个判别函数(因为 min{自变量个数，分组数 − 1} = min{4,3 − 1} = 2)。

结果 3-6-1 Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	3.842a	96.2	96.2	0.891
2	0.152a	3.8	100.0	0.363

a. First 2 canonical discriminant functions were used in the analysis.

结果 3-6-2 Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	0.179	89.355	6	0.000
2	0.868	7.339	2	0.025

结果 3-6-3 Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
Population in thousands	−0.006	1.012
Gross domestic product / capita	0.895	0.066
Predominant climate	0.486	−0.019

结果 3-6-4 Canonical Discriminant Function Coefficients

	Function	
	1	2
Population in thousands	0.000	0.000
Gross domestic product / capita	0.000	0.000
Predominant climate	0.333	−0.013
(Constant)	−3.723	−0.431

Unstandardized coefficients

结果 3-6-5 Structure Matrix

	Function	
	1	2
Gross domestic product / capita	0.875*	−0.090
People living in cities (%)a	0.587*	−0.157
Predominant climate	0.447*	0.177
Population in thousandsa	−0.048	0.998*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions

Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

a. This variable not used in the analysis.

结果 3-6-6 Functions at Group Centroids

Region or economic group	Function	
	1	2
OECD	2.415	−0.095
Pacific/Asia	−0.938	0.569
Africa	−1.879	−0.374

Unstandardized canonical discriminant functions evaluated at group means

结果 3-6-1 说明第一个判别函数已经包含了 96.2% 的信息，基本上通过第一个判别函数就能判定。当第一个判别函数不能完全确定时，可以用第二个判别函数来确定。

结果 3-6-1 说明第一个判别函数已经包含了 96.2% 的信息，基本上通过第一个判别函数就能判定。当第一个判别函数不能完全确定时，可以用第二个判别函数来确定。

结果 3-6-2 说明两个判别函数都是显著的。

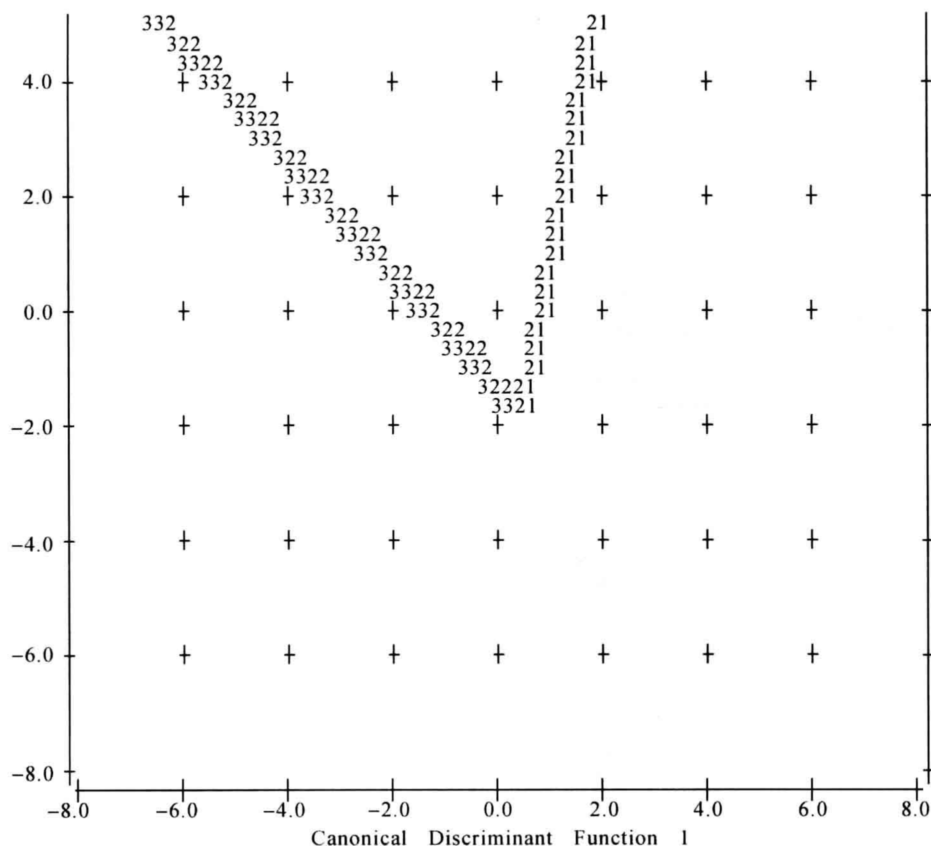
结果 3-6-3 给出了两个判别函数关于标准化以后的判别自变量的系数。

结果 3-6-4 给出了两个判别函数关于未标准化的判别自变量的系数。

结果 3-6-5 Structure Matrix 中给出的是判别得分关于判别自变量的相关系数，它表示相应自变量的判别能力。由于第一个判别函数占了 96.2% 的信息，而且自变量 gdp_cap 在第一个判别函数中的相关系数最大(0.875)，所以 gdp_cap 对判别结果的贡献最大。

结果 3-6-6 给出了各个组关于两个函数的判别重心。这个判别重心在下面的联合分布图和单独分布图中可以直观地理解。

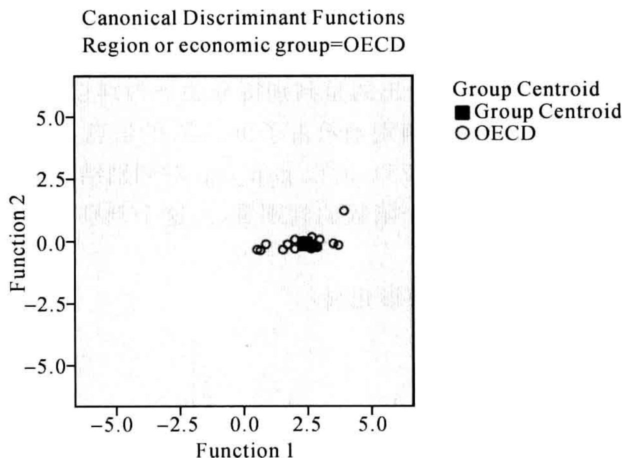
(4) 结果 3-7：逐步判别的结果：图形化显示



结果 3-7-1　领域图

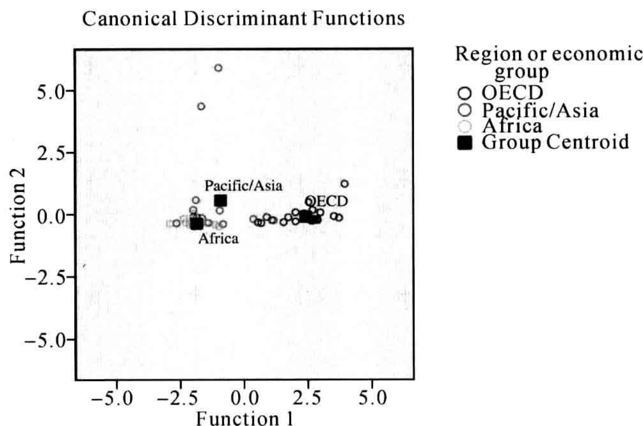
由于领域图太大，所以只是截取了其中一部分，具体地可以看自己 SPSS 程序的结果输出窗口。

领域图的横坐标是第一个判别函数的值，纵坐标是第二个判别函数的值，“Y”字型的边界构成了不同分组的界线。即当两个判别函数值构成的坐标落在上面领域图的左下角，那么这个样本被分在第三组；当两个判别函数值构成的坐标落在上面领域图的右下角，那么这个样本被分在第1组；当两个判别函数值构成的坐标落在上面领域图的上面部分，那么这个样本被分在第二组。



结果 3-7-2　单独分布图

上图举例地给出了第一组的单独分布图，第二、三组的单独分布图见 SPSS 程序的结果输出窗口。图中给出了所有用于分析的、并且被判在第一组的样本在（Function1，Function2）坐标系中的分布。并且也给出了这一组的重心。



结果 3-7-3　联合分布图

联合分布图只是将三组单独分布图放在一张图中，从中可以比较不同组的重心。

实验四　主成分分析

4.1　实验背景

由于变量间存在一定的相关关系，因此有可能用较少数的综合指标分别综合存在于各变量中的各类信息。本实验以 Midwestern 银行在 1969—1971 年之间雇员情况的数据，选取其中的五个变量作主成分分析。SPSS 默认保留特征根大于 1 的主成分，在本例中将看到保留 3 个主成分为宜，这 3 个主成分集中了原始 5 个变量中 90.66% 的信息。

需注意的是 SPSS 在调用 Factor Analyze 过程进行分析时，SPSS 会自动对原始数据进行标准化处理，所以在得到计算结果后的变量都是指经过标准化处理后的变量，但 SPSS 并不直接给出标准化后的数据，如需要得到标准化数据，则需调用 Descriptives 过程进行计算。

4.2　实验步骤和结果分析

(一) 实验数据

例 4.1　SPSS 自带的数据集 Employee data. sav 为 Midwestern 银行在 1969—1971 年之间雇员情况的数据，共包括 474 条观测及如下 10 个变量：Id(观测号)、Gender(性别)、Bdate(出生日期)、Educ(受教育程度(年数))、Jobcat(工作种类)、Salary(目前年薪)、Salbegin(开始受聘时的年薪)、Jobtime(受雇时间(月))、Prevexp(受雇以前的工作时间(月))、Minority(是否少数民族)。下面我们用主成分分析方法处理该数据，以期用少数变量来描述该地区居民的雇佣情况。

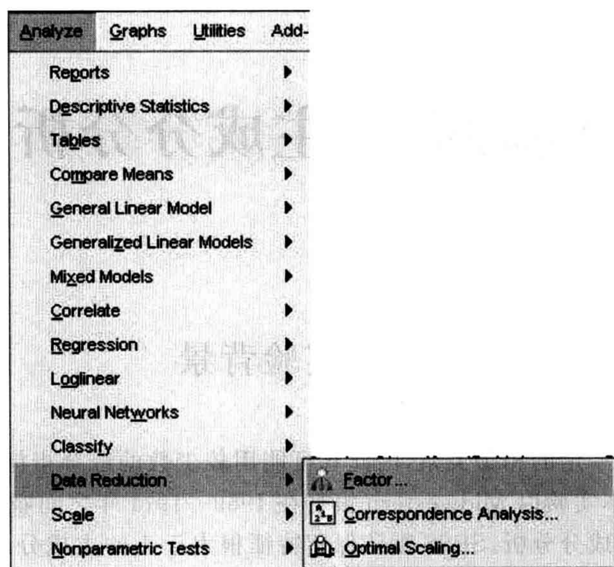
(二) 实验步骤

1. 打开数据

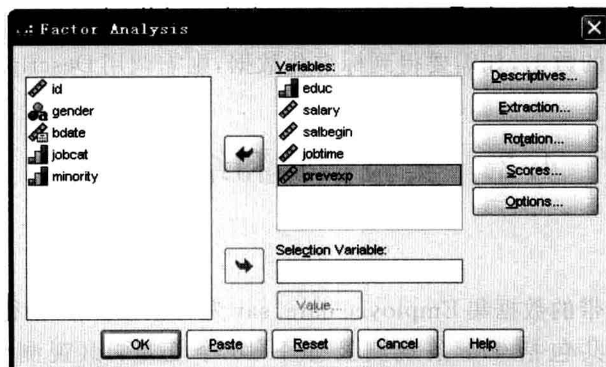
找到 SPSS 程序的安装目录，在其中的 Samples 文件夹中找到 Employee data. sav，双击打开，在打开成功后将数据文件另存为“主成分分析. sav”。

2. 主成分分析

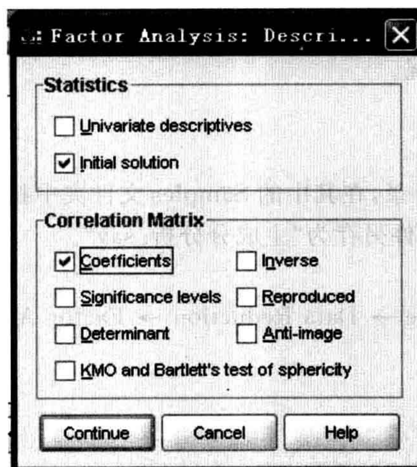
按如下方式选择：Analyze → Data Reduction → Factor Analysis：



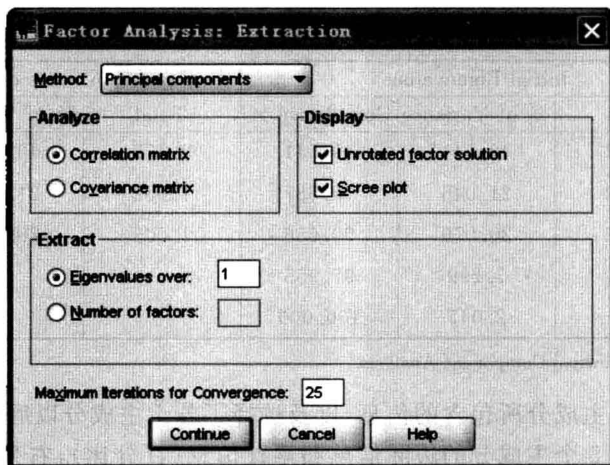
将“educ”、“salary”、“salbegin”、“jobtime”和“prevexp”这五个变量选入 Variable：



在 Descriptives 选项中选择 Correlation Matrix 下面的“Coefficients”，点击 Continue：



在 Extraction 选项中，选择 Display 选项下面的“Scree plot”，点击 Continue。然后点击 OK：



（三）结果分析

（1）结果 4-1：变量相关性

结果 4-1　Correlation Matrix

		Educational Level (years)	Current Salary	Beginning Salary	Months since Hire	Previous Experience (months)
Correlation	Educational Level (years)	1.000	0.661	0.633	0.047	−0.252
	Current Salary	0.661	1.000	0.880	0.084	−0.097
	Beginning Salary	0.633	0.880	1.000	−0.020	0.045
	Months since Hire	0.047	0.084	−0.020	1.000	0.003
	Previous Experience (months)	−0.252	−0.097	0.045	0.003	1.000

上表给出了五个变量之间的相关系数，从矩阵中可以看出还是存在比较大的相关性的，特别是第一、第二变量之间，第二、第三变量之间。所以对这五个变量进行主成分分析是有必要的。

（2）结果 4-2：主成分分析结果

结果 4-2-1　Communalities

	Initial	Extraction
Educational Level (years)	1.000	0.754
Current Salary	1.000	0.896
Beginning Salary	1.000	0.916
Months since Hire	1.000	0.999
Previous Experience (months)	1.000	0.968

Extraction Method: Principal Component Analysis.

上表给出了变量共同度，即给出了每个变量的原始信息在新的主成分中得以体现的比

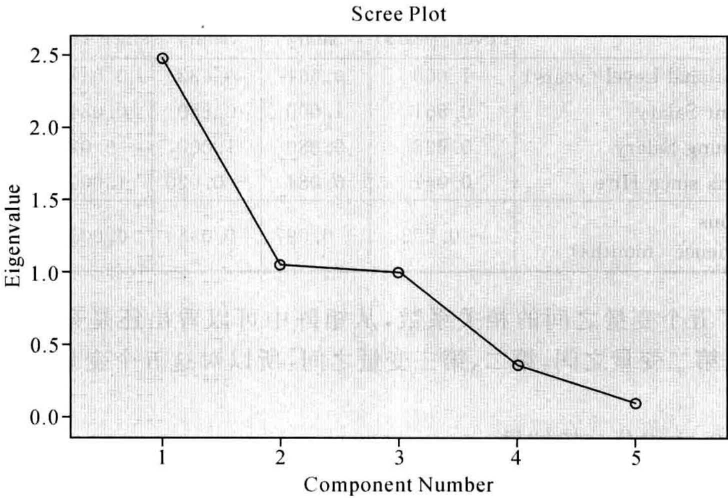
例，例如第一个变量在新成分中只保留了75.4%的信息，第二个变量的信息保留了89.6%，而且第三个变量保留的信息相对后面两个变量的信息也较少。这个和结果4-1中前三个变量之间存在相关性的判断是一致的。

结果 4-2-2 Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.477	49.541	49.541	2.477	49.541	49.541
2	1.052	21.046	70.587	1.052	21.046	70.587
3	1.003	20.070	90.656	1.003	20.070	90.656
4	.365	7.299	97.955			
5	.102	2.045	100.000			

Extraction Method：Principal Component Analysis.

上表给出了每个主成分所包含的信息，以及选择了若干主成分以后的累计信息。左列给出了从1个主成分到5个主成分的边际信息和累计信息（百分比），右列给出了抽出的主成分以及相关的信息。之所以抽出3个主成分，是因为我们在Extraction选项中默认了Extract "Eigenvalues over 1"，所以只抽出了信息（特征值）大于1的主成分。从最后一列可以看出，我们抽出的主成分保留了原始信息的90.656%。



结果 4-2-3　碎石图

上面的碎石图给出了特征值（主成分信息含量）的分布。由于第3个主成分的特征值还是大于1的，而第4个特征值急剧下降，并且远小于1，所以只保留前面3个主成分。

结果 4-2-4 Component Matrixa

	Component		
	1	2	3
Educational Level (years)	0.846	−0.194	−0.014
Current Salary	0.940	0.104	0.029
Beginning Salary	0.917	0.264	−0.077
Months since Hire	0.068	−0.052	0.996
Previous Experience (months)	−0.178	0.965	0.069

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

之前的判断说明选择了3个主成分。上表给出了这3个主成分通过原先5个变量(已标准化)线性组成的系数。"Component Matrix"是指初始因子载荷矩阵,每一个载荷量表示主成分与对应变量的相关系数。

结果4-2-4中的数据除以主成分相对应的特征值开平方根便得到两个主成分中每个指标所对应的系数。将第一列,第二列,第三列的数据分别除以:

$\sqrt{\lambda_1}=\sqrt{2.477}$,$\sqrt{\lambda_2}=\sqrt{1.052}$,$\sqrt{\lambda_3}=\sqrt{1.003}$ 可得系数矩阵:

结果 4-2-5 系数矩阵

	Component		
	1	2	3
Educational Level (years)	0.537536	−0.18914	−0.01398
Current Salary	0.597262	0.101397	0.028957
Beginning Salary	0.582648	0.257393	−0.07688
Months since Hire	0.043206	−0.0507	0.994509
Previous Experience (months)	−0.1131	0.940848	0.068897

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

根据上表,可得到三个主成分:

$$F_1=0.5375x_1+0.5972x_2+0.5826x_3+0.0432x_4-0.1131x_5$$

$$F_2=-0.1891x_1+0.1013x_2+0.2573x_3-0.0507x_4+0.9408x_5$$

$$F_3=-0.01398x_1+0.02895x_2-0.07688x_3-0.9945x_4+0.0688x_5$$

从系数的大小可以看出,第一个主成分对前三个变量的系数较大,所以可以把第一个主成分称为综合成分;第二个主成分的几乎所有质量都集中在第五个变量上,所以第二个主成分可以称为先前工作经验成分;第三个主成分的几乎所有质量都集中在第四个变量上,所以第三个主成分可以称为工龄成分。

(四) 另一个例子的主成分分析

例 4.2 本案例为例1.1中的数据,具体进行主成分分析的步骤和结论分析同例4.1,所以只列出结果。

(1) 结果4-3:变量相关性

结果 4-3　Correlation Matrix

		净资产收益率	总资产报酬率	资产负债率	总资产周转率	流动资产周转率	已获利息倍数	销售增长率	资本积累率
Correlation	净资产收益率	1.000	0.891	0.054	0.688	0.721	0.518	0.651	0.485
	总资产报酬率	0.891	1.000	−0.158	0.572	0.708	0.665	0.528	0.405
	资产负债率	0.054	−0.158	1.000	0.143	−0.043	−0.407	0.161	−0.281
	总资产周转率	0.688	0.572	0.143	1.000	0.782	0.142	0.547	0.342
	流动资产周转率	0.721	0.708	−0.043	0.782	1.000	0.272	0.452	0.389
	已获利息倍数	0.518	0.665	−0.407	0.142	0.272	1.000	0.228	0.458
	销售增长率	0.651	0.528	0.161	0.547	0.452	0.228	1.000	0.402
	资本积累率	0.485	0.405	−0.281	0.342	0.389	0.458	0.402	1.000

（2）结果 4-4：主成分分析结果

结果 4-4-1　Communalities

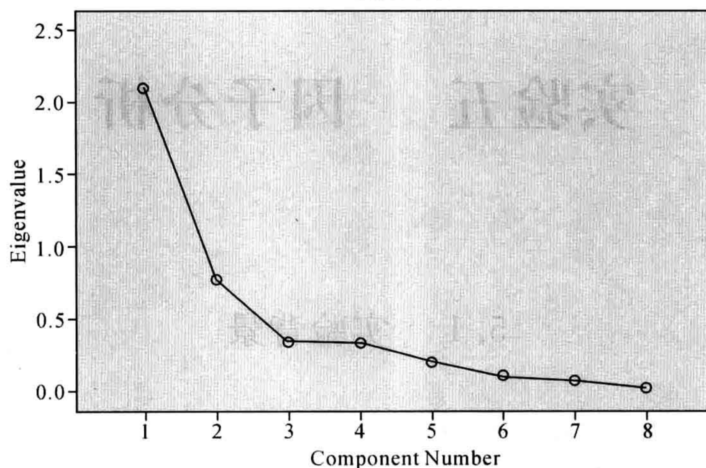
	Initial	Extraction
净资产收益率(%)	1.000	0.883
总资产报酬率(%)	1.000	0.830
资产负债率(%)	1.000	0.735
总资产周转率	1.000	0.764
流动资产周转率(%)	1.000	0.715
已获利息倍数	1.000	0.749
销售增长率(%)	1.000	0.585
资本积累率(%)	1.000	0.501

Extraction Method：Principal Component Analysis.

结果 4-4-2　Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.202	52.529	52.529	4.202	52.529	52.529
2	1.560	19.498	72.027	1.560	19.498	72.027
3	0.703	8.793	80.820			
4	0.683	8.540	89.361			
5	0.416	5.201	94.561			
6	0.219	2.737	97.298			
7	0.158	1.978	99.276			
8	0.058	0.724	100.000			

Extraction Method：Principal Component Analysis.



结果 4-4-3　碎石图

结果 4-4-4　Component Matrixa

	Component	
	1	2
净资产收益率(%)	0.934	0.105
总资产报酬率(%)	0.903	−0.124
资产负债率(%)	−0.100	0.851
总资产周转率	0.771	0.411
流动资产周转率(%)	0.825	0.188
已获利息倍数	0.594	−0.630
销售增长率(%)	0.702	0.302
资本积累率(%)	0.620	−0.342

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

实验五　因子分析

5.1　实验背景

因子分析是研究如何以最少的信息丢失将众多原有变量浓缩成少数几个因子，如何使因子具有一定的命名解释性的多元统计分析方法。因子分析具体的步骤有：

（一）判断观测数据是否适合作因子分析。

原始变量个数较少，可对相关矩阵进行检验，如果相关矩阵中的大部分相关系数小于0.3，则不适合做因子分析；当原始变量个数较多时，经常采用的方法为巴特利特球体检验和KMO。Bartlett 球体检验的目的是检验相关矩阵是否是单位矩阵；KMO 是 Kaiser-Meyer-Olkin 的取样适当性量数。

（二）抽取共同因子，确定因子的数目和求因子解的方法。

本实验采用“主成分分析法”(principal components analysis) 为决定因素抽取的方法。因子数目借助两个准则来确定：一是特征值(eigenvalue) 准则，二是碎石图检验(scree test) 准则。

（三）因子旋转：使因子更具有命名可解释性。

通常最初因素抽取后，对因素无法作有效的解释。这时往往需要进行因子旋转(rotation)，通过坐标变换使因子解的意义更容易解释。这里我们将选择最大变异法(Varimax)。

（四）计算因子得分

本步骤正是通过各种方法计算各样本在各因子上的得分，为进一步的分析奠定基础。

本实验用例 5.1 和例 5.2 两个例子说明利用 SPSS 的 Factor Analysis 模块进行因子分析的方法。例 5.1 是 SPSS 自带的 Employee data. sav 数据集；例 5.2 为例 1.1 中的数据。

5.2　实验步骤和结果分析

（一）实验数据

例 5.1　SPSS自带的数据集 Employee data. sav 为 Midwestern 银行在 1969—1971 年之间雇员情况的数据，共包括 474 条观测及如下 10 个变量：Id(观测号)、Gender(性别)、Bdate(出生日期)、Educ(受教育程度(年数))、Jobcat(工作种类)、Salary(目前年薪)、Salbegin(开始受聘时的年薪)、Jobtime(受雇时间(月))、Prevexp(受雇以前的工作时间

(月))、Minority(是否少数民族)。下面我们用因子分析方法处理该数据。

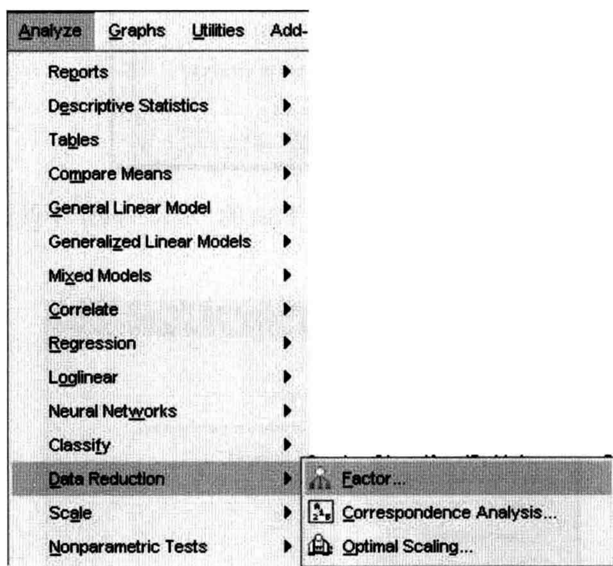
(二) 实验步骤

1. 打开数据:

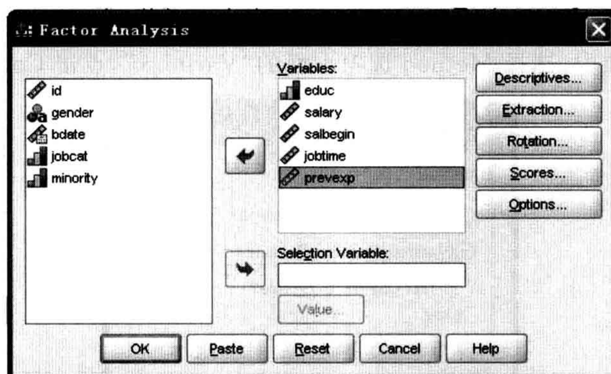
找到 SPSS 程序的安装目录,在其中的 Samples 文件夹中找到 Employee data. sav,双击打开,在打开成功后将数据文件另存为“因子分析. sav”。

2. 因子分析

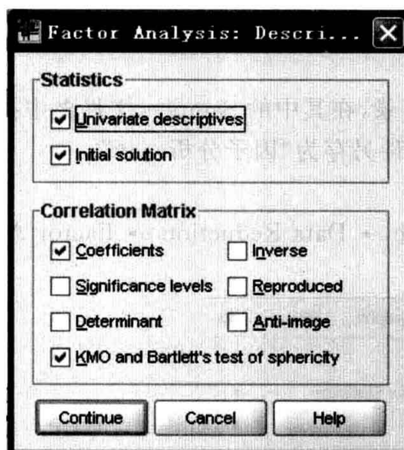
按如下方式选择:Analyze → Data Reduction → Factor Analysis:



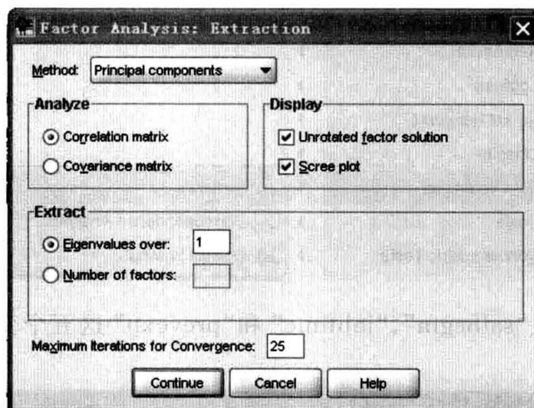
将“educ”、“salary”、“salbegin”、“jobtime” 和“prevexp” 这五个变量选入 Variable:



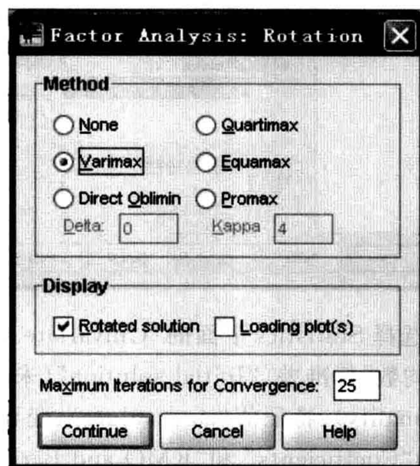
在 Descriptives 选项中,选择 Statistics 下面的“Univariate descriptives”(单变量描述性统计量):显示每一题项的平均数、标准差。“Initial solution”(未转轴之统计量):显示因素分析未转轴前之共同性(communality)、特征值(eigenvalues)、变异数百分比及累积百分比;选择 Correlation Matrix 下面的“Coefficients” 和“KMO and Bartlett's test of sphericity”,点击 Continue:



在 Extraction 选项中，选择 Display 选项下面的“Scree plot”，点击 Continue。然后点击 OK：



在 Rotation 选项中，选择 Method 选项下面的“Varimax”，点击 Continue：



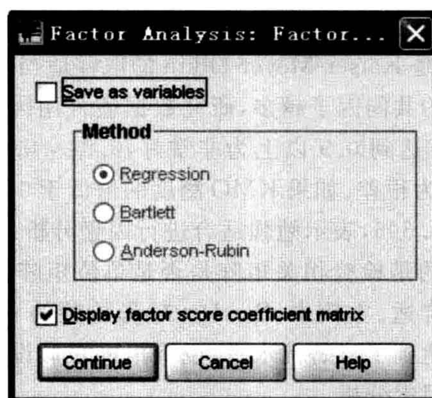
在 Scores 选项中，Save as variable 框勾选时可将新建立的因素分数存储至数据文件中，并产生新的变量名称(内定为 fact_1、fact_2 等)。在"Method"框中表示计算因素分数的方法有三种：

(1)Regression：使用回归法；

(2)Bartlett：使用 Bartlette 法；

(3)Anderson-Robin：使用 Anderson-Robin 法；

Display factor score coefficient matrix"(显示因素分数系数矩阵)选项勾选时可显示因素分数系数矩阵。本实验选择如下图，点击 Continue，点击 OK：



(三)结果分析

(1) 结果 5-1：变量相关性估计和检验

结果 5-1-1　Descriptive Statistics

	Mean	Std. Deviation	Analysis N
Educational Level (years)	13.49	2.885	474
Current Salary	$ 34,419.57	$ 17,075.661	474
Beginning Salary	$ 17,016.09	$ 7,870.638	474
Months since Hire	81.11	10.061	474
Previous Experience (months)	95.86	104.586	474

结果 5-1-2　Correlation Matrix

		Educational Level (years)	Current Salary	Beginning Salary	Months since Hire	Previous Experience (months)
Correlation	Educational Level (years)	1.000	0.661	0.633	0.047	−0.252
	Current Salary	0.661	1.000	0.880	0.084	−0.097
	Beginning Salary	0.633	0.880	1.000	−0.020	0.045
	Months since Hire	0.047	0.084	−0.020	1.000	0.003
	Previous Experience (months)	−0.252	−0.097	0.045	0.003	1.000

结果 5-1-3　KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.606
Bartlett's Test of Sphericity	Approx. Chi-Square	1094.808
	df	10
	Sig.	0.000

结果 5-1-1 给出了每个变量的均值方法估计。

结果 5-1-2 给出了五个变量之间的相关系数，如果相关矩阵中的大部分相关系数小于 0.3，则不适合作因子分析。从本例的矩阵中可以看出还是存在比较大的相关性的，特别是第一、第二变量之间，第二、第三变量之间。所以对这五个变量进行因子分析是有必要的。

结果 5-1-3 分析：KMO 是 Kaiser-Meyer-Olkin 的取样适当性量数。KMO 测度的值越高（接近 1.0 时），表明变量间的共同因子越多，研究数据适合用因子分析。通常按以下标准解释该指标值的大小：KMO 值达到 0.9 以上为非常好，0.8 ～ 0.9 为好，0.7 ～ 0.8 为一般，0.6 ～ 0.7 为差，0.5 ～ 0.6 为很差。如果 KMO 测度的值低于 0.5 时，表明样本偏小，需要扩大样本，此处的 KMO 值为 0.606，表示勉强适合进行因素分析。

Bartlett 球体检验的目的是检验相关矩阵是否是单位矩阵（identity matrix），如果是单位矩阵，则认为因子模型不合适。本例中，Bartlett 球形检验的 χ^2 值为 1094.808（自由度为 10），p 值为 0.000 < 0.01，达到了显著性水平，说明拒绝零假设而接受备择假设，即相关矩阵不是单位矩阵，适合进行因素分析。

（2）结果 5-2：主成分分析部分的结果：

结果 5-2-1　Communalities

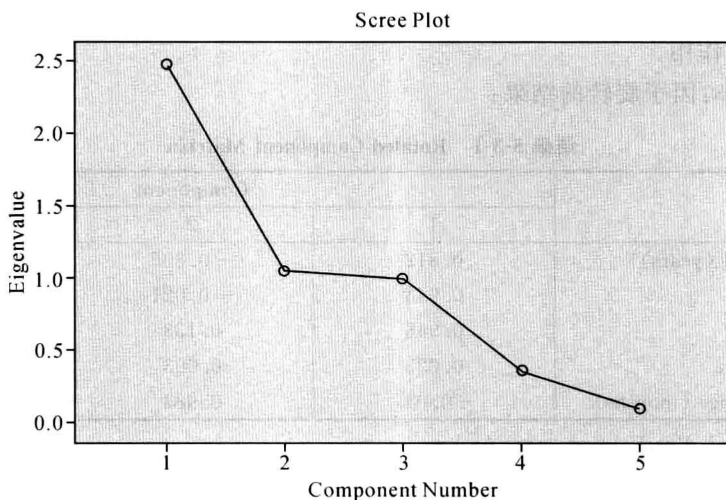
	Initial	Extraction
Educational Level (years)	1.000	0.754
Current Salary	1.000	0.896
Beginning Salary	1.000	0.916
Months since Hire	1.000	0.999
Previous Experience (months)	1.000	0.968

Extraction Method: Principal Component Analysis.

结果 5-2-2　Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.477	49.541	49.541	2.477	49.541	49.541
2	1.052	21.046	70.587	1.052	21.046	70.587
3	1.003	20.070	90.656	1.003	20.070	90.656
4	0.365	7.299	97.955			
5	0.102	2.045	100.000			

Extraction Method: Principal Component Analysis.



结果 5-2-4　碎石图

结果 5-2-4　Component Matrixa

	Component		
	1	2	3
Educational Level (years)	0.846	－0.194	－0.014
Current Salary	0.940	0.104	0.029
Beginning Salary	0.917	0.264	－0.077
Months since Hire	0.068	－0.052	0.996
Previous Experience (months)	－0.178	0.965	0.069

Extraction Method：Principal Component Analysis.

a. 3 components extracted.

结果 5-2-1 给出了变量共同度。表中显示各因子解释掉方差的比例，也称变量的共同度(Communality)。共同度从 0 到 1，0 为因子不解释任何方差，1 为所有方差均被因子解释掉。一个因子越大地解释掉变量的方差，说明因子包含原有变量信息的量越多。数值小就说明该变量不适合作因子，可在分析中将其排除在外。本实验除 Educational Level 的共同度小于 0.8 外(接近)，其余均大于 0.8，故五个变量适合因子分析。

结果 5-2-2 上表给出了每个主成分所包含的信息。上表中第一列为特征值(主成分的方差)，第二列为各个主成分的贡献率，第三列为累积贡献率，由上表看出前 3 个主成分的累计贡献率就达到了 90.656％ ＞ 85％，所以选取主成分个数为 3。

结果 5-2-3 上面的碎石图给出了特征值(主成分信息含量)的分布。由上图看出，成分数为 3 时，特征值的变化曲线趋于平缓，所以由碎石图也可大致确定出主成分个数为 3。与按累计贡献率确定的主成分个数是一致的。

结果 5-2-4 是因子载荷矩阵，是用标准化后的主成分近似表示标准化原始变量的系数矩阵，用 F1，F2，F3 表示各公因子，以 Beginning Salary 为例，即有：

Beginning Salary ≈ 0.917F1 ＋ 0.264F2 － 0.077F3

这 3 个主成分质量的侧重点已经比较明显了。所以其实对于这个例子来说已经没有必

要再进行因子旋转，所以这里体现不出因子分析的目的，后面例 5.2 的那个案例分析将充分体现因子旋转的作用。

（3）结果 5-3：因子旋转的结果：

结果 5-3-1 Rotated Component Matrixa

	Component		
	1	2	3
Educational Level (years)	0.812	−0.306	0.036
Current Salary	0.944	−0.021	0.066
Beginning Salary	0.946	0.133	−0.050
Months since Hire	0.023	0.003	0.999
Previous Experience (months)	−0.047	0.983	0.004

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 4 iterations.

结果 5-3-2 Component Transformation Matrix

Component	1	2	3
1	0.990	−0.134	0.046
2	0.137	0.989	−0.058
3	−0.038	0.064	0.997

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

结果 5-3-3 Component Score Coefficient Matrix

	Component		
	1	2	3
Educational Level (years)	0.314	−0.229	0.013
Current Salary	0.388	0.049	0.040
Beginning Salary	0.403	0.193	−0.074
Months since Hire	−0.017	0.011	0.994
Previous Experience (months)	0.051	0.921	0.012

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

结果 5-3-4 Component Score Covariance Matrix

Component	1	2	3
1	1.000	0.000	0.000
2	0.000	1.000	0.000
3	0.000	0.000	1.000

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

结果 5-3-1 给出了经过因子旋转以后的主成分系数。它所体现的在几个变量上质量的侧重与旋转之前没有本质上区别，事实上区别不是很大。

结果 5-3-2 给出了用于因子旋转的变换矩阵。从中可以发现这个矩阵和单位矩阵其实相

差不是很大，大部分质量集中在对角线上，其他地方的值都不是很大。这也印证了上面所说的本例没有太大必要进行因子旋转。

结果 5-3-3 给出了公因子用 5 个原始自变量表示的线性估计。

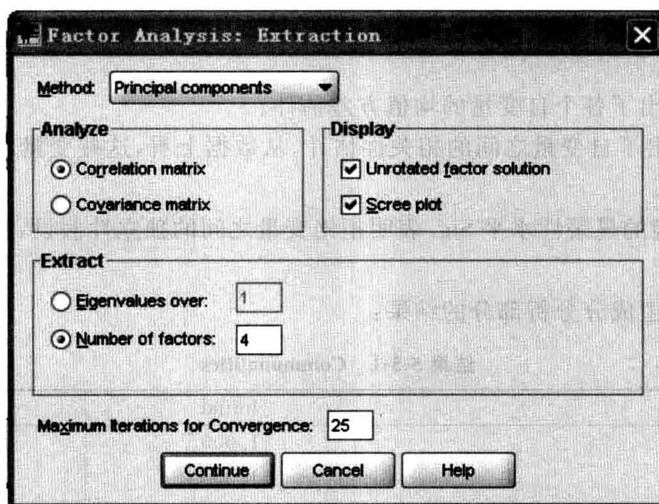
结果 5-3-4 给出了公因子个协方差矩阵。

(四) 另一个例子的因子分析

例 5.2 本案例为例 1.1 中的数据，具体进行因子分析的步骤同例 5.1，所以只列出结果并进行结果分析。

从上一次实验的第二个主成分分析的结论知道，如果只抽出特征值大于 1 的主成分，那么只能抽出 2 个主成分，而且抽出的总的信息只有 72.027%。为了能够从中抽出更多的信息，不妨选择抽出 4 个主成分。具体在操作上的变化为：

在 Extraction 选项中，选择 Extract 选项下面的"Number of factor"为 4，点击 Continue。然后点击 OK：



(1) 结果 5-4：变量相关性估计和检验

结果 5-4-1 Descriptive Statistics

	Mean	Std. Deviation	Analysis N
净资产收益率(%)	9.6872	6.70821	35
总资产报酬率(%)	7.2406	4.36864	35
资产负债率(%)	47.1163	15.19990	35
总资产周转率	0.4266	0.45635	35
流动资产周转率(%)	0.7717	0.68101	35
已获利息倍数	8.3086	8.42327	35
销售增长率(%)	17.9963	38.03200	35
资本积累率(%)	14.6774	21.86696	35

结果 5-4-2 Correlation Matrix

		净资产收益率	总资产报酬率	资产负债率	总资产周转率	流动资产周转率	已获利息倍数	销售增长率	资本积累率
Correlation	净资产收益率	1.000	0.891	0.054	0.688	0.721	0.518	0.651	0.485
	总资产报酬率	0.891	1.000	−0.158	0.572	0.708	0.665	0.528	0.405
	资产负债率	0.054	−0.158	1.000	0.143	−0.043	−0.407	0.161	−0.281
	总资产周转率	0.688	0.572	0.143	1.000	0.782	0.142	0.547	0.342
	流动资产周转率	0.721	0.708	−0.043	0.782	1.000	0.272	0.452	0.389
	已获利息倍数	0.518	0.665	−0.407	0.142	0.272	1.000	0.228	0.458
	销售增长率	0.651	0.528	0.161	0.547	0.452	0.228	1.000	0.402
	资本积累率	0.485	0.405	−0.281	0.342	0.389	0.458	0.402	1.000

结果 5-4-3 KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.737
Bartlett's Test of Sphericity	Approx. Chi−Square	181.204
	df	28
	Sig.	0.000

结果 5-4-1 给出了各个自变量的均值方差估计。

结果 5-4-2 给出了自变量之间的相关性估计。从数据上看，这些变量之间存在则非常严重的多重共线性性。

结果 5-4-2 检验的显著性水平 Sig. 表明拒绝变量之间的独立性假设，即他们之间存在相关性。

（2）结果 5-5：主成分分析部分的结果：

结果 5-5-1 Communalities

	Initial	Extraction
净资产收益率(%)	1.000	0.919
总资产报酬率(%)	1.000	0.943
资产负债率(%)	1.000	0.878
总资产周转率	1.000	0.881
流动资产周转率(%)	1.000	0.911
已获利息倍数	1.000	0.908
销售增长率(%)	1.000	0.799
资本积累率(%)	1.000	0.910

Extraction Method: Principal Component Analysis.

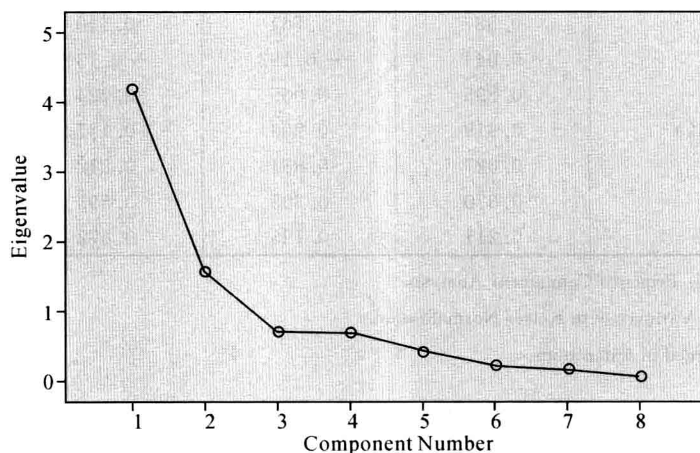
结果 5-5-2 Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.202	52.529	52.529	4.202	52.529	52.529	2.582	32.278	32.278
2	1.560	19.498	72.027	1.560	19.498	72.027	1.950	24.375	56.653
3	0.703	8.793	80.820	0.703	8.793	80.820	1.351	16.891	73.545

续表

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
4	0.683	8.540	89.361	0.683	8.540	89.361	1.265	15.816	89.361
5	0.416	5.201	94.561						
6	0.219	2.737	97.298						
7	0.158	1.978	99.276						
8	0.058	0.724	100.000						

Extraction Method: Principal Component Analysis.



结果 5-5-3　碎石图

结果 5-5-4　Component Matrixa

	Component			
	1	2	3	4
净资产收益率(%)	0.934	0.105	0.001	0.189
总资产报酬率(%)	0.903	−0.124	−0.175	0.287
资产负债率(%)	−0.100	0.851	0.223	0.306
总资产周转率	0.771	0.411	−0.209	−0.270
流动资产周转率(%)	0.825	0.188	−0.374	−0.237
已获利息倍数	0.594	−0.630	0.052	0.395
销售增长率(%)	0.702	0.302	0.461	0.042
资本积累率(%)	0.620	−0.342	0.474	−0.429

Extraction Method: Principal Component Analysis.

a. 4 components extracted.

结果 5-5-1，从每个变量抽出信息的百分比可以看出，几乎所有的变量被抽出的信息都达到或超过了 80%。这说明当选择抽出 4 个主成分时，抽出的信息比 2 个主成分多了，并且已经足够。

结果5-5-2给出了抽出的各个主成分的信息比例，以及因子旋转以后各个主成分的信息比例。从中知道：虽然总体4个主成分的信息没有变，但是4个主成分间的信息相差变小了。

结果5-5-3给出了碎石图。

结果5-5-4给出了没有经过因子旋转的主成分系数，各个主成分没有非常明显地将质量侧重在某几个变量上，所以很难对各个主成分进行解释。

（3）结果5-6：因子旋转的结果。

结果 5-6-1　Rotated Component Matrixa

	Component			
	1	2	3	4
净资产收益率(%)	0.640	0.615	0.299	0.203
总资产报酬率(%)	0.587	0.761	0.140	−0.008
资产负债率(%)	0.044	−0.193	−0.131	0.907
总资产周转率	0.895	0.065	0.224	0.161
流动资产周转率(%)	0.919	0.209	0.137	−0.066
已获利息倍数	0.027	0.881	0.212	−0.294
销售增长率(%)	0.370	0.305	0.593	0.467
资本积累率(%)	0.211	0.195	0.872	−0.260

Extraction Method：Principal Component Analysis.

Rotation Method：Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

结果 5-6-2　Component Transformation Matrix

Component	1	2	3	4
1	0.708	0.562	0.425	0.036
2	0.381	−0.421	−0.149	0.809
3	−0.499	0.016	0.776	0.386
4	−0.322	0.712	−0.441	0.441

Extraction Method：Principal Component Analysis.

Rotation Method：Varimax with Kaiser Normalization.

结果 5-6-3　Component Score Coefficient Matrix

	Component			
	1	2	3	4
净资产收益率(%)	0.093	0.294	−0.036	0.186
总资产报酬率(%)	0.110	0.450	−0.275	0.033
资产负债率(%)	−0.111	0.081	−0.043	0.761
总资产周转率	0.506	−0.294	−0.017	−0.069
流动资产周转率(%)	0.562	−0.196	−0.194	−0.253
已获利息倍数	−0.277	0.662	−0.078	−0.038
销售增长率(%)	−0.154	0.067	0.523	0.443
资本积累率(%)	−0.113	−0.261	0.896	−0.189

Extraction Method：Principal Component Analysis.

Rotation Method：Varimax with Kaiser Normalization.

结果 5-6-4 Component Score Covariance Matrix

Component	1	2	3	4
1	1.000	0.000	0.000	0.000
2	0.000	1.000	0.000	0.000
3	0.000	0.000	1.000	0.000
4	0.000	0.000	0.000	1.000

Extraction Method：Principal Component Analysis.

Rotation Method：Varimax with Kaiser Normalization.

结果 5-6-1，上面因子旋转以后的因子载荷矩阵可以知道：

第一个公因子主要的载荷集中在总资产周转率与流动资产周转率上，同时也有相对比较多的载荷在净资产收益率与总资产报酬率上，因此第一个因子体现公式的总体运营情况，称为综合因子。

第二个因子主要的载荷集中在已获利息倍数、总资产报酬率和净资产收益率上，可以理解成公司的获利能力因子。

第三个因子主要的载荷集中在资本累计率和销售增长率上，因此这个因子

可以理解成公司的发展能力因子。

第四个因子主要的载荷集中在资产负债率上，所以它是公司的负债因子。

结果 5-6-2 图给出了用于因子旋转的变换矩阵。此时这个变换矩阵与单位矩阵相差很大，这说明因子转换对因子的表达式变化很大。

结果 5-6-3 给出了公因子用 5 个原始自变量表示的线性估计。

结果 5-6-4 给出了公因子个协方差矩阵。

实验六　典型相关分析

6.1　实验背景

在 SPSS 中可以用两种方法来拟合典型相关分析，第一种是采用 Manova 过程来拟合，第二种是采用专门提供的宏程序来拟合，第二种方法在使用上非常的简单，而输出的结果非常的详细。因此这里只对第二种方法进行介绍。

利用 SPSS 软件对 C. R. Rao(1952) 关于典型相关的经典例子进行分析。表 6-1 列举了 25 个家庭的成年长子和次子的头长和头宽。利用典型相关分析法分析长子和次子头型的相关性。

6.2　实验步骤和结果分析

(一) 实验数据

为了研究兄长的头型与弟弟的头型间的关系，研究者随机抽查了 25 个家庭的两兄弟的头长和头宽，资料见表 6-1，希望求得两组变量的典型变量及典型相关系数。这里，代表兄长头型的变量为第一组变量，代表弟弟头型的变量为第二组变量，希望求得的是两组变量间的相关性。

表 6-1　兄弟头长与头宽的相关资料　　(单位:mm)

序号	X_1 兄头长	X_2 兄头宽	Y_1 弟头长	Y_2 弟头宽
1	191.00	155.00	179.00	145.00
2	183.00	153.00	188.00	149.00
3	189.00	150.00	190.00	149.00
4	192.00	150.00	187.00	151.00
5	174.00	150.00	185.00	152.00
6	163.00	137.00	161.00	130.00
7	181.00	145.00	182.00	146.00
8	174.00	143.00	178.00	147.00
9	190.00	163.00	187.00	150.00
10	195.00	149.00	201.00	152.00
11	176.00	144.00	171.00	142.00
12	197.00	159.00	189.00	152.00

续表

序号	X_1 兄头长	X_2 兄头宽	Y_1 弟头长	Y_2 弟头宽
13	179.00	158.00	186.00	148.00
14	190.00	159.00	195.00	157.00
15	195.00	155.00	183.00	158.00
16	175.00	140.00	165.00	137.00
17	176.00	139.00	176.00	143.00
18	181.00	148.00	185.00	149.00
19	208.00	157.00	192.00	152.00
20	188.00	152.00	197.00	159.00
21	183.00	147.00	174.00	147.00
22	188.00	151.00	187.00	158.00
23	186.00	153.00	173.00	148.00
24	192.00	154.00	185.00	152.00
25	197.00	167.00	200.00	158.00

(二)实验步骤

1. 点击“Files → New → Syntax”,打开 SPSS 的语法输入窗口,如图 6-1。



图 6-1 进入 SPSS 的语法输入窗口

2. 在语法输入窗口输入如下程序,调查典型相关分析的专用模块,具体程序见图 6-2:

输入时要注意“Canonical correlation. sps”程序所在的根目录,注意变量组的格式和空格。这里,“Canonical correlation. sps”所在的根目录为:D:\Program Files\SPSSInc\SPSS16\Samples,变量组 1 包含 X1,X2,变量组 2 包含 Y1,Y2。

3. 在图 6-2 所示的窗口中,按照 Run 菜单 —> ALL 开始运行典型相关计算程序,主要显示结果如下。

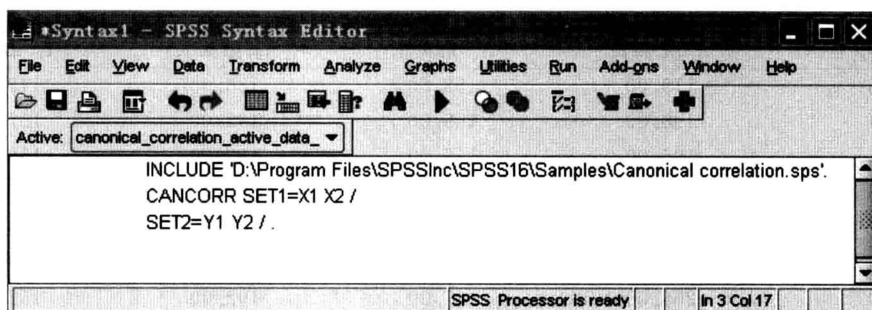


图 6-2　SPSS 的语法输入窗口程序图

(三) 实验结果与分析

1. 实验输出结果

表 6-2　变量内部相关矩阵表

Correlations for Set-1

	X1	X2
X1	1.0000	.7346
X2	.7346	1.0000

Correlations for Set-2

	Y1	Y2
Y1	1.0000	.8393
Y2	.8393	1.0000

表 6-3　变量间的两两相关矩阵

Correlations Between Set-1 and Set-2

	Y1	Y2
X1	.7108	.7040
X2	.6932	.7086

表 6-4　典型相关系数表

Canonical Correlations

1	.789
2	.054

表 6-5　相关系数检验表

Test that remaining correlations are zero:

	Wilk's	Chi-SQ	DF	Sig.
1	.377	20.964	4.000	.000
2	.997	.062	1.000	.803

表 6-6　典型变量 1 系数列表

Standardized Canonical Coefficients for Set-1

	1	2
X1	-.552	-1.366
X2	-.522	1.378

Raw Canonical Coefficients for Set-1

	1	2
X1	-.057	-.140
X2	-.071	.187

表 6-7　典型变量 2 系数列表

Standardized Canonical Coefficients for Set-2

	1	2
Y1	-.504	-1.769
Y2	-.538	1.759

Raw Canonical Coefficients for Set-2

	1	2
Y1	-.050	-.176
Y2	-.080	.262

表 6-8　第一变量中的典型相关系数表

Canonical Loadings for Set-1

	1	2
X1	-.935	-.354
X2	-.927	.375

Cross Loadings for Set-1

	1	2
X1	-.737	-.019
X2	-.731	.020

表 6-9　第二变量中的典型相关系数表

```
Canonical Loadings for Set-2
            1          2
Y1      -.956      -.293
Y2      -.962       .274


Cross Loadings for Set-2
            1          2
Y1      -.754      -.016
Y2      -.758       .015
```

表 6-10　冗余度分析表 1

```
        Redundancy Analysis:
Proportion of Variance of Set-1 Explained by Its Own Can. Var.
           Prop Var
CV1-1          .867
CV1-2          .133
```

表 6-11　冗余度分析表 2

```
Proportion of Variance of Set-1 Explained by Opposite Can.Var.
           Prop Var
CV2-1          .539
CV2-2          .000
```

表 6-12　冗余度分析表 3

```
Proportion of Variance of Set-2 Explained by Its Own Can. Var.
           Prop Var
CV2-1          .920
CV2-2          .080
Proportion of Variance of Set-2 Explained by Opposite Can. Var.
           Prop Var
CV1-1          .572
CV1-2          .000

------ END MATRIX -----
```

表 6-13 补充说明表

```
The canonical scores have been written to the active file.
Also, a file containing an SPSS Scoring program has been written
To use this file GET a system file with the SAME variables
Which were used in the present analysis. Then use an INCLUDE command
to run the scoring program.
For example :
GET FILE anotherfilename
INCLUDE FILE "CC__.INC".
EXECUTE.
```

2. 实验结果分析

(1) 首先给出的是两组变量内部各自的相关矩阵(如表 6-2),第一组变量间的相关系数为0.7436,即兄长的头长和头宽的相关系数为0.7436。第二组变量间的相关系数为0.8393。可见兄弟间头长和头宽是有相关性。

(2) 表 6-3 给出的是两组变量间各变量的两两相关矩阵,从此表中可以看出相关系数均在 0.7 左右,可见兄弟的指标间确实存在相关性,这里需要做的就是提取出综合指标来代表这种相关性。

(3) 表 6-4 给出的是提取的两个典型相关系数的大小,可见第一典型相关系数为 0.789,第二典型相关系数为 0.054。

(4) 表 6-5 为检验各典型相关系数有无统计意义的。第一对典型变量显著性检验的 χ^2 统计量为20.964,p 值为0,说明第一对典型相关变量显著相关。第二对典型相关变量显著性检验的 χ^2 统计量为 0.062,p 值为 0.803,说明第二对典型相关变量相关性不显著,可见第一对典型相关系数有统计学意义,而第二对典型相关系数则没有统计学意义。因此,只取第一对典型相关变量即可。

(5) 表 6-6 显示各典型变量与变量组 1 中各变量间标化与未标化的系数列表,由此可以写出典型变量的转换公式(标化的) 为:

$L_1 = 0.552 * X_1 + 0.552 * X_2$ $\qquad$ $L_2 = 1.336 * X_1 - 1.378 * X_2$

(6) 表 6-7 显示各典型变量变量组 2 中各变量间标化与未标化的系数列表,同上可以写出典型变量的转换公式(标化的) 为:

$M_1 = 0.504 * Y_1 + 0.538 * Y_2$ $\qquad$ $M_2 = 1.769 * Y_1 - 1.759 * Y_2$

(8) 表 6-8 显示第一变量组中各变量分别与自身、相关的典型变量的相关系数,可见它们主要和第一对典型变量的关系比较密切。

(9) 表 6-9 显示第二变量组中各变量分别与自身、相关的典型变量的相关系数,可见它们主要和第一对典型变量的关系比较密切。

下面即将输出的是冗余度(Redundancy) 分析结果,它列出各典型相关系数所能解释员变量变异的比例,可以用来辅助判断需要保留多少个典型相关系数。

(10) 表 6-10 显示的是第一组变量的变异可被自身的典型变量所解释的比例,可见第一典型变量解释了总变异的 86.7 %,而第二典型变来能够则只解释了总变异的 13.3 %。

(11) 表 6-11 显示的是第一组变量的变异能被它们相对的典型变量所解释的比例,可见

第二典型变量的解释度非常小。

(12) 表 6-12 显示的是第二变量组的变异分别能被自身、相对的典型变量所解释的比例,可见结论和上面一样,第二对变量的贡献非常小。因此综合上述冗余分析结果,我们只需要保留第一对典型变量即可。

(13) 最后系统给出说明:标化变量已被写入当前文件,同时相应的计算程序也以文件形式被存储在当前目录中,可以使用 GET 命令进入数据文件,再使用 INCLUDE 命令来调用相应程序。

实验七　对应分析

7.1　实验背景

对应分析法的整个处理过程由两部分组成:列联表和关联图。在关联图上,我们能把众多的样品和众多的变量同时作到同一张图解上,将样品的大类及其属性在图上直观而又明了地表示出来,具有直观性。

本实验通过例 7.1,例 7.2 说明如何运用 SPSS 软件的 Correspondence Analysis 模块进行对应分析。

7.2　实验步骤和结果分析

(一) 实验数据

例 7.1　选用 SPSS 软件自带的 GSS93 subset. sav 数据,该数据在 SPSS 软件的安装目录下可以找到,该数据共包括 1500 个观测,67 个变量。选用该数据集中 Degree(学历) 与 Race(人种) 变量为例来进行他们之间的对应分析。其中 Degree 变量各个取值的含义如下:0— 中学以下(less than high school),1— 中学(high school),2— 专科(junior college),3— 本科(bachelor),4— 研究生(graduate),7,8,9— 缺失;Race 变量各个取值的含义如下:1— 白种人(white),2— 黑种人(black),3— 其他(other)。

(二) 实验步骤

1. 打开数据

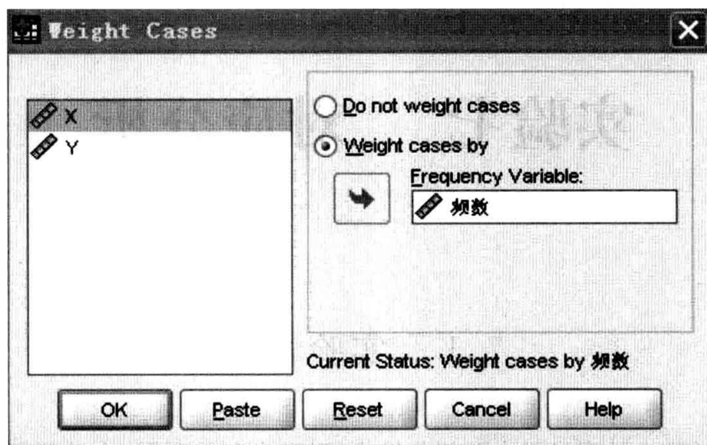
找到 SPSS 程序的安装目录,在其中的 Samples 文件夹中找到 GSS93 subset. sav,双击打开,在打开成功后将数据文件另存为“7.1 对应分析. sav”。

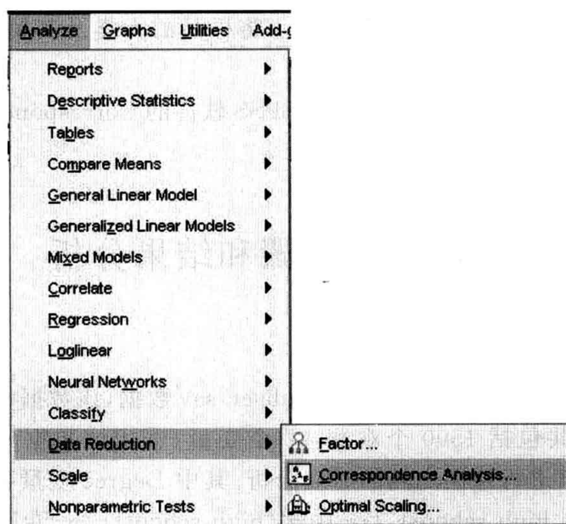
2. 对数据文件进行处理

如果数据文件中有频数变量,应该首先做加权处理。在 Data → Weight Cases,如下图把频数选择进右边框。例 7.1 人数并无加权,一条记录信息代表一个人,因此本实验的这个步骤省略。

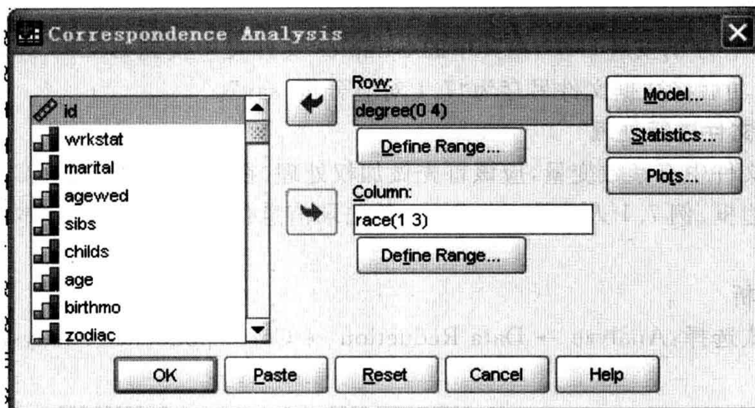
3. 对应分析

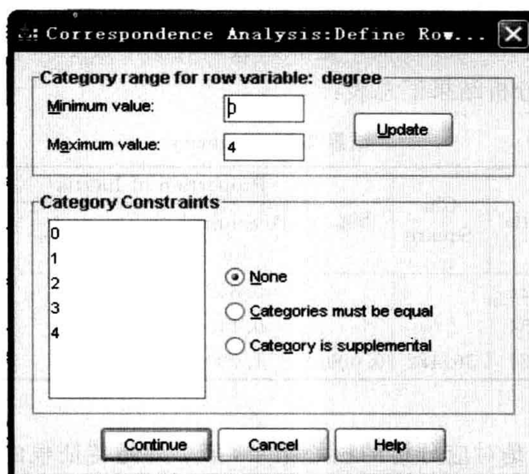
按如下方式选择:Analyze → Data Reduction → Correspondence Analysis:



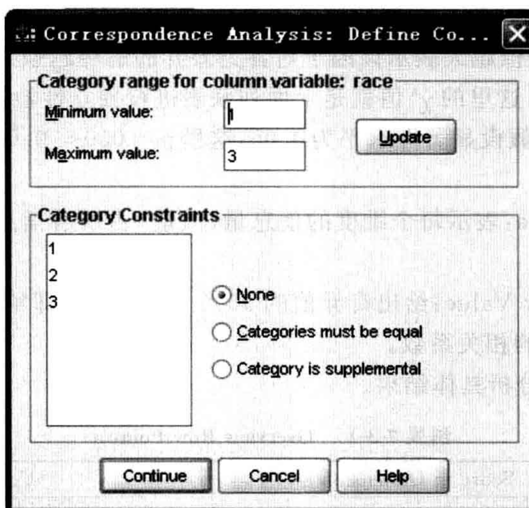


将"degree"选入 Row，点击它下面的 Define Range，在"Minimum value"这一项中输入 0，在"Maximum value"这一项中输入 4，点击 Update。然后点击 Continue：下图中的"id"应为"degree"。





将“race”选入Column，点击它下面的Define Range，在“Minimum value”这一项中输入1，在“Maximum value”这一项中输入3，点击Update。然后点击Continue。然后点击OK：



（三）结果分析

（1）结果7-1：对应分析表（列联表）。

结果 7-1　Correspondence Table

R's Highest Degree	Race of Respondent			
	white	black	other	Active Margin
Less than HS	214	48	17	279
High school	658	92	30	780
Junior college	74	13	3	90
Bachelor	209	7	18	234
Graduate	99	7	7	113
Active Margin	1254	167	75	1496

结果 7-1 给出了由原始数据按 Degree 与 Race 分类的列联表，由于原始数据中有 4 条记录有缺失，所以观测总数 n = 1496 而不是原始数据观测个数 1500。

（2）结果 7-2：对应分析结果汇总表：

结果 7-2　Summary

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation
								2
1	0.144	0.021			0.852	0.852	0.021	0.065
2	0.060	0.004			0.148	1.000	0.026	
Total		0.024	36.482	0.000a	1.000	1.000		

a. 8 degrees of freedom

结果 7-2：Summary 是对应分析的核心结果，第 1 列是特征根的编号，由对应分析的原理，提取的特征根数为：min{行变量类别数，列变量类别数}-1，此处正好为 min{5,3} − 1 = 2。

Singular Value & Inertia：惯量（Inertia）相当于因子分析中的特征根，奇异值（Singular Value）就是惯量的平方根。第 1 个特征值的值最大，第 2 个特征值较小，类似于因子分析中特征值的含义知道，特征值越大表示该因子对各类差异的解释越强。

Chi-Square & Sig.：这里的 χ^2 值就是上面列联表进行独立性检验所用到的 χ^2 检验统计量和显著性水平（p 值），假设显著性水平为 0.05，这里 p = 000 < 0.05，可认为行变量与列变量有显著的相关关系。

Proportion of Inertia：表示每个维度的信息量（惯量）占所有信息量的比例，第一维度可解释总信量的 85.2%。

Confidence Singular Value：给出奇异值的 95% 可信区间，即给出了一个标准差的值。同时还给出了两个维度的相关系数。

（3）结果 7-3：对应分析具体结果。

结果 7-3-1　Overview Row Pointsa

R's Highest Degree	Mass	Score in Dimension		Inertia	Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
		1	2		1	2	1	2	Total
Less than HS	0.186	−0.462	−0.414	0.008	0.276	0.531	0.750	0.250	1.000
High school	0.521	−0.078	0.192	0.002	0.022	0.322	0.285	0.715	1.000
Junior college	0.060	−0.304	0.193	0.001	0.039	0.037	0.857	0.143	1.000
Bachelor	0.156	0.723	−0.203	0.012	0.566	0.107	0.968	0.032	1.000
Graduate	0.076	0.429	−0.041	0.002	0.096	0.002	0.996	0.004	1.000
Active Total	1.000			0.024	1.000	1.000			

a. Symmetrical normalization

上表称为行点汇总图，给出了行变量（Degree）各类别的分析结果概况。具体各项的解释如下：

Mass：表示每个类别的样本数占总样本数的比例；

Score in Dimension：表示行变量各分类对于第 1 和第 2 个因子上的因子载荷，同时它们也将成为散点图中相应类别的坐标。

Inertia 给出了每个类别的特征值。

Contribution：Of Point to Inertia of Dimension：第 6 和第 7 列是行变量各分类对第 1 和第 2 个因子的差异影响程度。可以看出，Bachelor 对第一个因子值影响的差异最大，达到 56.6%。

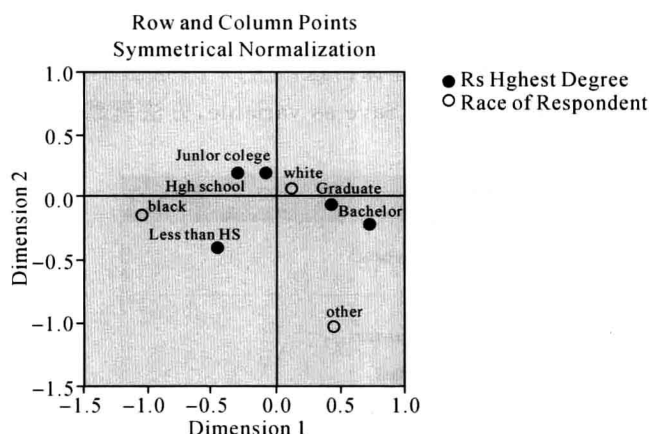
Of Dimension to Inertia of Point 下面表示两个因子对行变量各分类差异的解释程度，例如：Less than HS，第一个因子解释了 75%，第二个因子解释了 25%，两因子共解释了 100%，损失信息 0。

结果 7-3-2　Overview Column Pointsa

Race of Respondent	Mass	Score in Dimension		Inertia	Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
		1	2		1	2	1	2	Total
white	0.838	0.113	0.079	0.002	0.074	0.088	0.830	0.170	1.000
black	0.112	−1.051	−0.134	0.018	0.855	0.033	0.993	0.007	1.000
other	0.050	0.452	−1.026	0.005	0.071	0.879	0.318	0.682	1.000
Active Total	1.000			0.024	1.000	1.000			

a. Symmetrical normalization

上表称为列点汇总图，给出了行变量（Race）各类别的分析结果概况。具体各项的解释同上面行点汇总图。



结果 7-3-3　对应分析图（编辑过）

上面的图称为对应分析图，对应分析所有的主要信息和结论都可以在这个上面体现。由于 SPSS 输出的原图对两个变量标示的颜色区别不大，所以将上图中行变量（Degree）由空心环涂成了实心圆，并且加上了过原点的坐标系。

对应分析图主要用来考察不同变量的类别之间的联系。判断的规则为：

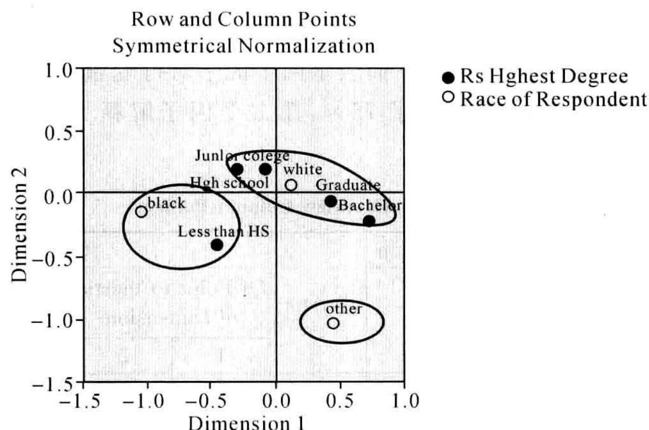
1. 落在从图形原点（0,0）出发相同方位上大致相同区域内的不同变量的分类点彼此有

联系；

2. 散点间距离越近，说明关联倾向越明显；

3. 散点离原点越远，也说明关联倾向越明显。

对于本例，从上面的对应分析图可以以及判断方法可以得到以下结论：



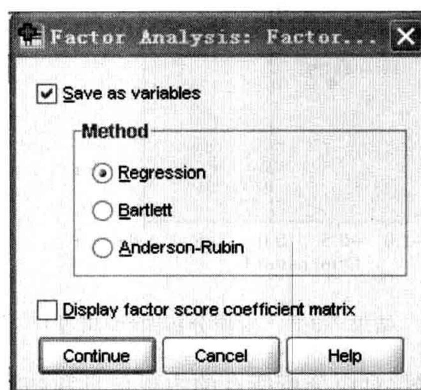
1. 黑种人(Black)与低于高中学历(Less than HS)有较强的关系，即黑人的教育水平一般在高中以下；

2. 白种人(White)与高中以上的四种学历的距离都比较近；

3. 其他人种(other)没有特别明显的特征。

(四) 另一个例子的因子分析

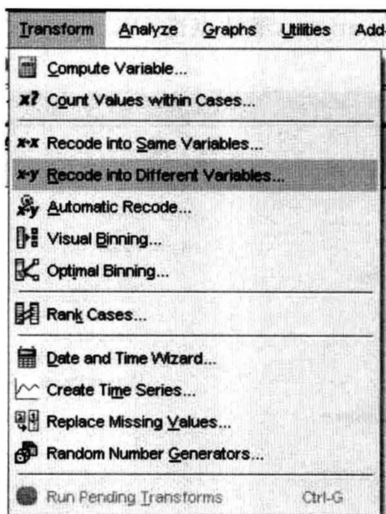
例 7.2 本案例为例 1.1 中的数据。在实验六中对这个数据进行过因子分析，并且当时设定抽出的因子数为 4 个，但是并没有选择保留这四个因子。这里不妨再快速地作一次因子分析，区别只是在 Scores 这个选项中选择 Save as variable，方法就默认为“Regression”：



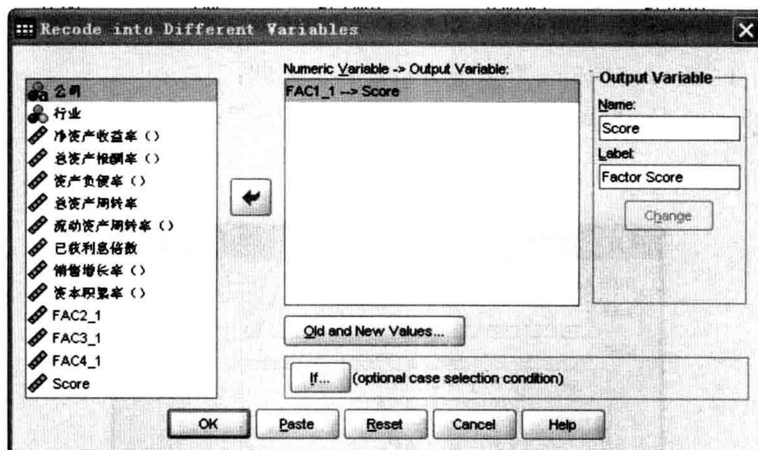
通过这样做了因子分析以后可以发现，在数据窗口中多了四个变量：Fac1_1、Fac2_1、Fac3_1 和 Fac4_1。这里所需要用到的数据并不是这四个连续型变量，而是希望按照 Fac1_1 的取值将所有公司关于 Fac1_1 取值的大小划分为 4 个等级：

Fac1_1 >= 0.5；　0 <= Fac1_1 < 0.5；　−0.5 <= Fac1_1 < 0；　Fac1_1 <−0.5；

并且分别将它们标示为 1、2、3 和 4。这些对于数据的初步处理可以按照如下方式进行：按如下方式选择：Transform → Recode into Different Variables：



将 FAC1_1 选入右边 Numeric Variable → Output Variable 方框内；点击方框内的"FAC1_1"变量，最右边的 Output Variable 方框被激活，依次在 Name 和 Label 下面输入"Score"和"Factor Score"：



点击 Old and New Values 选项，跳出下面的对话框。依次按下面方式计算新变量的值：

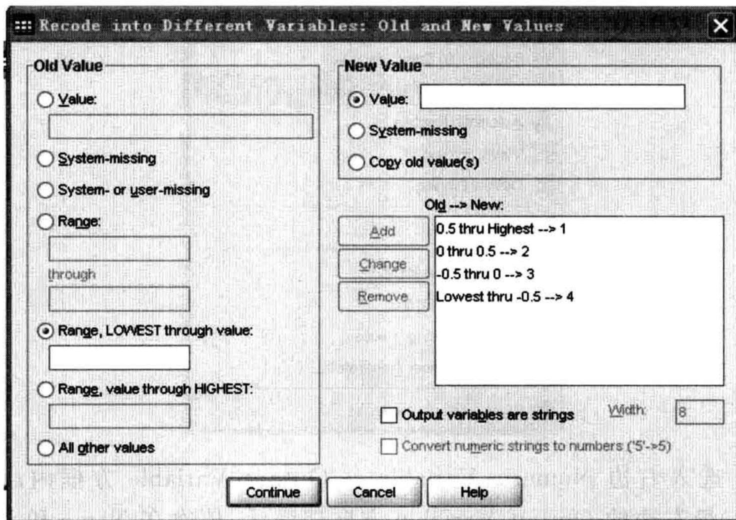
1）选中 Range，value though HIGHEST，然后在下面输入的方框内输入 0.5，在右侧 New Value 下面的 Value 方框内输入 1，点击 Add。这一步就把 Fac1_1 取值在[0.5，+∞）内的值都转化为 1；

2）选中 Range，分别在下面的两个方框内输入 0 和 0.5，然后在右侧 New Value 下面的 Value 方框内输入 2，点击 Add。这一步就把 Fac1_1 取值在[0，0.5）内的值都转化为 2；

3）选中 Range，分别在下面的两个方框内输入 −0.5 和 0，然后在右侧 New Value 下面

的 Value 方框内输入 3，点击 Add。这一步就把 Fac1_1 取值在[−0.5,0)内的值都转化为 3；

4）选中 Range，LOWEST through value，然后在下面输入的方框内输入 −0.5，在右侧 New Value 下面的 Value 方框内输入 4，点击 Add。这一步就把 Fac1_1 取值在(−∞，−0.5)内的值都转化为 4；然后点击 Continue，继续点击 OK。

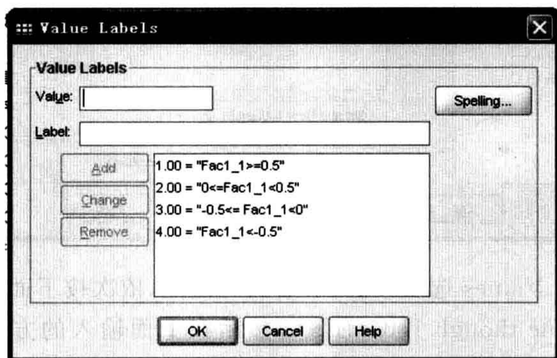


完成了上面的转换，在数据窗口就可以发现多了一个变量“Score”，它就是每个样本的 Fac1_1 评分等级。为了更加直观地显示，可以在 Variable View 更改变量“Score”每个数值对应的含义，即添加如下信息：

1 = “Fac1_1 >= 0.5”；2 = “0 <= Fac1_1 < 0.5”；

3 = “−0.5 <= Fac1_1 < 0”；4 = “Fac1_1 <−0.5”；

见下图：



具体的操作方法之前介绍过，这里就不作详细说明了。

下面对这个案例正式地进行对应分析。具体的分析步骤和案例 7.1 一样，这里不再具体写出，只分析对应分析的结果：

(1) 结果 7-4:对应分析表(列联表)。

结果 7-4 Correspondence Table

行业	Factor Score				
	Fac1_1 >= 0.5	0 <= Fac1_1 < 0.5	−0.5 <= Fac1_1 < 0	Fac1_1 < −0.5	Active Margin
电力煤气及水的生产和供应业	3	1	5	2	11
房地产业	0	1	5	9	15
信息技术业	4	3	1	1	9
Active Margin	7	5	11	12	35

上表给出的是对应分析表,即列联表。

(2) 结果 7-5:对应分析结果汇总表。

结果 7-5 Summary

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation 2
1	0.621	0.385			0.837	0.837	0.116	0.081
2	0.274	0.075			0.163	1.000	0.158	
Total		0.460	16.105	0.013a	1.000	1.000		

a. 6 degrees of freedom

上表给出了整个对应分析的结果汇总表。

(3) 结果 7-6:对应分析具体结果。

结果 7-6-1

行业	Mass	Score in Dimension		Inertia	Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
		1	2		1	2	1	2	Total
电力煤气及水的生产和供应业	0.314	−0.202	0.761	0.058	0.021	0.665	0.137	0.863	1.000
房地产业	0.429	0.818	−0.264	0.186	0.462	0.109	0.956	0.044	1.000
信息技术业	0.257	−1.117	−0.490	0.216	0.517	0.226	0.922	0.078	1.000
Active Total	1.000			0.460	1.000	1.000			

a. Symmetrical normalization

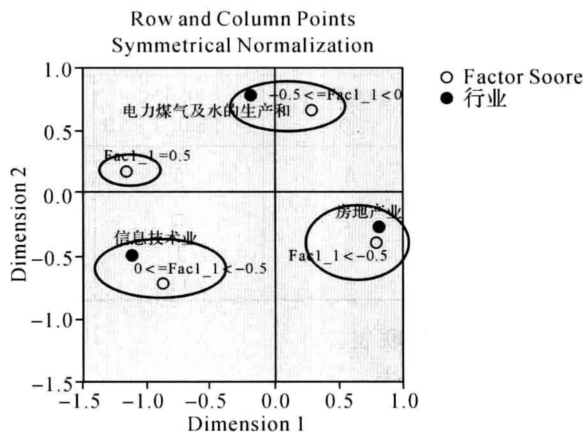
上表称为行点汇总表。

结果 7-6-2

Factor Score	Mass	Score in Dimension		Inertia	Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
		1	2		1	2	1	2	Total
Fac1_1 >= 0.5	0.200	−1.168	0.169	0.171	0.439	0.021	0.991	0.009	1.000
0 <= Fac1_1 < 0.5	0.143	−0.881	−0.711	0.089	0.179	0.264	0.777	0.223	1.000
−0.5 <= Fac1_1 < 0	0.314	0.288	0.663	0.054	0.042	0.505	0.300	0.700	1.000
Fac1_1 < −0.5	0.343	0.784	−0.410	0.147	0.340	0.210	0.893	0.107	1.000
Active Total	1.000			0.460	1.000	1.000			

a. Symmetrical normalization

上表称为列点汇总图。



结果 7-6-3　对应分析图(编辑过)

上图为对应分析图,对上述的解释为:

1. 信息技术业的综合能力是三个行业里面最好的,它的综合因子得分平均在[0,0.5)内;

2. 房地产业的综合能力最差,综合因子得分平均落在小于－0.5;

3. 电力煤气及水的生产和供应业的综合能力介于前两个行业之间,综合因子得分在[－0.5,0)内。

实验八　多维标度分析

8.1　实验背景

多维标度法解决的问题是：当 n 个对象(object) 中各对对象之间的相似性(或距离) 给定时，确定这些对象在低维空间中的表示(感知图 Perceptual Mapping)，并使其尽可能与原先的相似性(或距离)“大体匹配”，使得由降维所引起的任何变形达到最小。

以 SPSS 自带文件 World95. sav 为例，对亚洲国家和地区的 17 个国家的人口寿命情况进行分析。选择以下变量：urban(城市人口比例)，lifeexpf(女性平均寿命)，lifeexpm(男性平均寿命)，gdp_cap(人均 GDP)，death_rt(千人死亡率)，birth_rt(千人出生率)，literacy(受教育人口比例) 所涉及的统计数据，对 17 个国家进行多维标度分析。

8.2　实验步骤和结果分析

(一) 实验数据

以 SPSS 自带文件 World95. sav 为例，对亚洲国家和地区的 17 个国家的人口寿命情况进行分析，在 Data → Select case 对话框的 If 过滤条件中输入过滤条件“region = 3”，得到 17 个国家和地区的数据。

(二) 实验步骤

1. 主菜单中选择 Analyze → Scale → Multidimensional Scaling (ALSCAL)，就进入多维标度法的主对话框(图 8-1)。在左上方的变量列表选择以下变量：urban(城市人口比例)，lifeexpf(女性平均寿命)，lifeexpm(男性平均寿命)，gdp_cap(人均 GDP)，death_rt(千人死亡率)，birth_rt(千人出生率)，literacy(受教育人口比例)。由于原始数据不是距离阵，因此需要在下方 Distances 单选项中选择 Create distances from data，这时 Measure 子对话框被激活，默认计算 Euclidean distance，即欧氏距离。

2. 点击进入 Measure 子对话框，对距离阵进行设定(图 8-2)。由于我们的变量都是连续数值型的，所以应在 Measure 单选项中选择 Interval。并在其下方的 Transform Values 栏中选择变量标准化变换的方式，这里我们选择 Z scores 和 By variable，表示对变量进行正态标准化。然后在 Create Distance Matrix 单选项中选择 Between cases，表示计算样品之间的距离阵。设置完毕后，点击 Continue 回到主对话框。

3. 在主对话框中点击进入 Model 子对话框，如图 8-3。这里可以设定变量取值的类型。

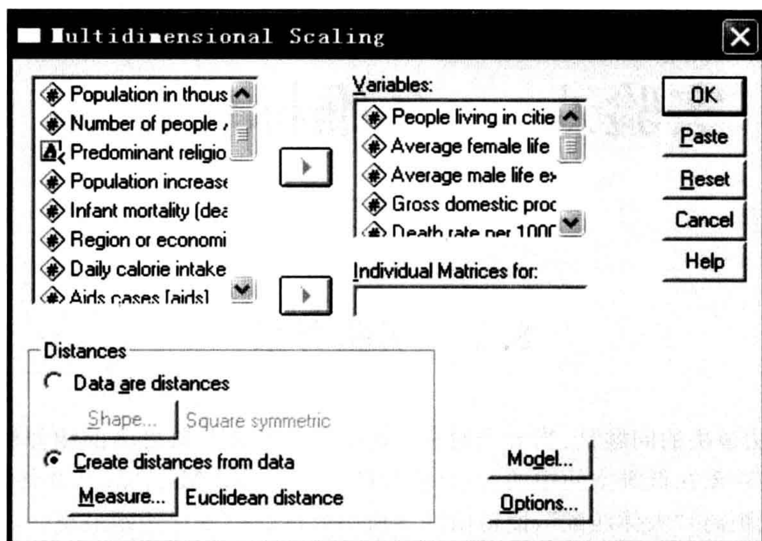


图 8-1　多维标度法的主对话框

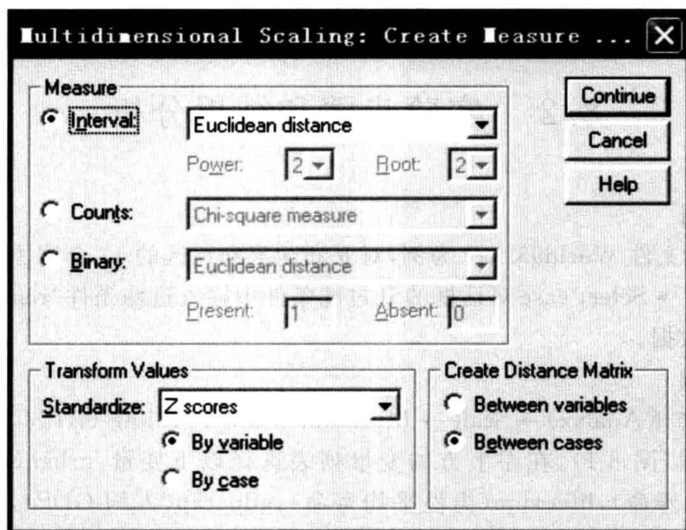


图 8-2　Measure 子对话框

在 Level of Measurement 中选择 Interval，即连续取值的数值型变量。其他设置无需改变，点击 Continue 返回主对话框。

4. 点击进入 Options 子对话框（图 8-4），该对话框中提供了一些结果显示的选择。Display 栏中默认不输出任何图表。选择 Group plots 项可得到多维标度图，这里图表的维度由 Model 中的 Dimensions 中填入最小维度 Minimum 和最大维度 Maximum 决定；选择 Data matrix 项可得到距离阵和拟合构造点的坐标；而 Model and options summary 是显示出多维

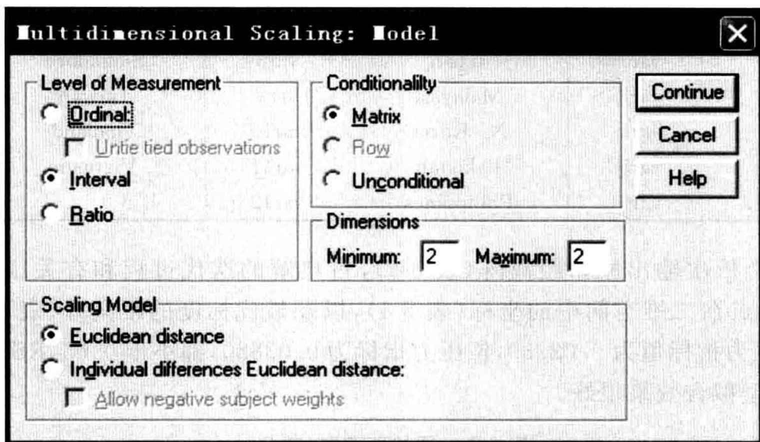


图 8-3 Model 子对话框

标度法中的参数设置，计算方法等。这里我们选择 Group plots 和 Data matrix 项后，点击 Continue 返回主对话框，再点击 OK 运行。

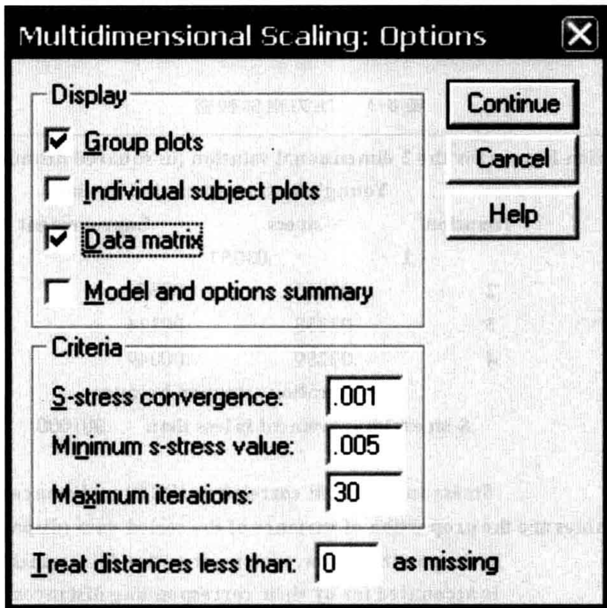


图 8-4 Options 子对话框

(三) 实验结果与分析

1. 样品验证表，发现有一个样品存在缺失值。查原始数据后发现 Taiwan 缺少千人死亡率，该样品被去除。国家地区的编号如下(表 8-1)。

表 8-1　国家和地区的编号

Afghanistan	var1	Indonesia	var7	S. Korea	var13
Bangladesh	var2	Japan	var8	Singapore	var14
Cambodia	var3	Malaysia	var9	Taiwan	*
China	var4	N. Korea	var10	Thailand	var15
Hong Kong	var5	Pakistan	var11	Vietnam	var16
India	var6	Philippines	var12		

2. SPSS 会依次输出原始距离阵(表 8-2),古典解的迭代过程和有关压力指标值(表 8-3),拟合构造点在二维空间中的坐标(表 8-4),以及最优标度的距离阵(表 8-5)。在表 8-3 中,Young 氏压力指标值为 0.02289,K 压力指标为 0.03880,都小于 0.05。RSQ = 0.99485。这些都说明模型拟合效果很好。

表 8-2　原始距离阵(部分)

Raw (unscaled) Data for Subject 1					
1	2	3	4	5	
1	0				
2	3.15	0			
3	1.794	1.451	0		
4	5.822	3.144	4.177	0	
5	7.905	5.685	6.554	3.59	0

表 8-3　压力指标检验

Iteration history for the 2 dimensional solution (in squared distances)
Young's S-stress formula 1 is used.

Iteration	S-stress	Improvement
1	.03057	
2	.02463	.00594
3	.02338	.00124
4	.02289	.00049

Iterations stopped because
S-stress improvement is less than .001000

Stress and squared correlation (RSQ) in distances
RSQ values are the proportion of variance of the scaled data (disparities)
in the partition (row, matrix, or entire data) which
is accounted for by their corresponding distances.
Stress values are Kruskal's stress formula 1.

For matrix
Stress = .03880　RSQ = .99485

表 8-4 拟合点的在 2 维标度中的坐标(部分)

Configuration derived in 2 dimensions Stimulus Coordinates			
Stimulus Number	Stimulus Name	Dimension	
		1	2
1	VAR1	2.8077	−0.7825
2	VAR2	1.4351	0.0200
3	VAR3	1.9799	−0.2425
4	VAR4	−0.1950	0.5249
5	VAR5	−1.7190	−0.7151

表 8-5 最优标度的距离阵(部分)

Optimally scaled data (disparities) for subject 1					
	1	2	3	4	5
1	0.000				
2	1.676	0.000			
3	0.856	0.648	0.000		
4	3.293	1.673	2.298	0.000	
5	4.553	3.210	3.736	1.942	0.000

3. 接下来是欧氏距离下的 16 个国家和地区的拟合构造点的二维图(图 8-5),从图上可以看出比较发达的地区基本都在第三个象限,如香港,日本,新加坡。而中国和泰国,菲律宾等国较为接近。而线性拟合散点图(图 8-6) 从图形上告诉我们采用欧氏距离来拟合原始数据的距离阵是非常合适的。

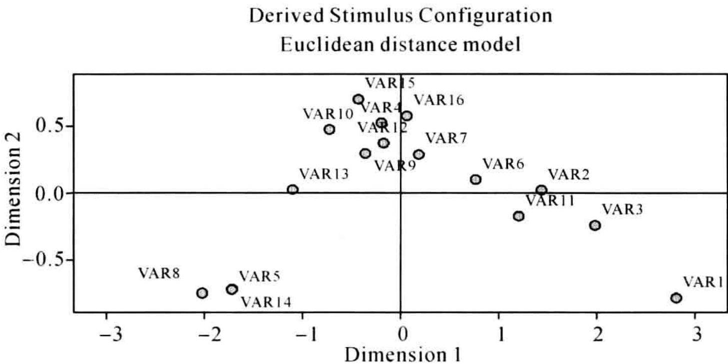


图 8-5 拟合构造点的二维坐标图

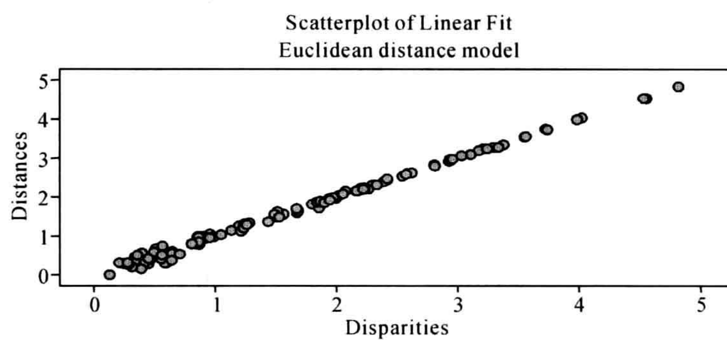


图 8-6 欧氏距离下的线性拟合散点图

参考文献

[1] 于秀林,任雪松.多元统计分析[M].北京:中国统计出版社,1998.

[2] 朱道元.多元统计分析与软件 SAS[M].南京:东南大学出版社,1999。

[3] R. A. Johnson, D. W. Wichern. 实用多元统计分析[M].北京:清华大学出版社,2001。

[4] 何晓群.多元统计分析[M].北京:中国人民大学出版社,2004.

[5] 高惠璇.应用多元统计分析[M].北京:北京大学出版社,2005.

[6] 张润楚.多元统计分析[M].北京:科学出版社,2006.

[7] 王静龙.多元统计分析[M].北京:科学出版社,2008.

[8] 王学仁,王松桂.实用多元统计分析[M].上海:上海科技出版社,1990.

[9] 汪冬华.多元统计分析与 SPSS 应用[M].上海:华东理工大学出版社,2010.

[10] 安德森.多元统计分析导论[M].张润楚译.北京:人民邮电出版社,2010.

[11] 哈德勒.应用多元统计分析[M].陈诗一译.北京:北京大学出版社,2011.

[12] 方开泰.实用多元统计分析[M].上海:华东师范大学出版社,1989.

[13] 王瑞庆.二维联合正态分布伪随机数生成算法的研究与实现[J],电脑学习,2008(2),9—11.

[14] 殷菲,潘晓平,吴震.Chernoff 脸谱图的改进[J],中国卫生统计,2003,8(4),194—196.

[15] 程正东,章毓晋,樊祥,朱斌.常用 Fisher 判别函数的判别矩阵研究[J],自动化学报,2010,36(10),1361—1370.

[General Information]
书名=多元统计概论与实验
作者=刘桂梅，林伟然编著
页数=183
SS号=13366256
DX号=
出版日期=2013.08
出版社=浙江大学出版社

封面
书名
版权
前言
目录
第一篇　多元统计分析原与方法
　　第一章　绪论
　　　　1.1多元统计简介
　　　　1.2主要内容安排
　　第二章　多元数据图表示法
　　　　2.1散点图矩阵
　　　　2.2雷达图
　　　　2.3调和曲线图
　　　　2.4脸谱图
　　第三章　均值向量和协方差阵的检验
　　　　3.1随机向量
　　　　3.2多元正态分布
　　　　　　3.2.1多元正态分布的定义及基本性质
　　　　　　3.2.2多元正态分布的参数估计
　　　　3.3均值向量的检验
　　　　　　3.3.1单个正态总体Np （μ，∑）均值向量的检验
　　　　　　3.3.2两个正态总体Np （μ1，∑1）和Np（μ2，∑2）均值向量的检验
　　　　　　3.3.3多个正态总体均值向量的检验（多元方差分析）
　　　　3.4协差阵的检验
　　　　　　3.4.1一个正态总体协方差阵的检验
　　　　　　3.4.2多个正态总体协方差阵的检验
　　第四章　聚类分析
　　　　4.1距离
　　　　　　4.1.1聚类数据的标准化处理
　　　　　　4.1.2样品距离的定义
　　　　4.2系统聚类法
　　　　　　4.2.1类间的距离
　　　　　　4.2.2四种系统聚类法
　　　　4.3 K-均值聚类法
　　第五章　判别分析
　　　　5.1判别分析简介
　　　　5.2距离判别法
　　　　　　5.2.1两组距离判别
　　　　　　5.2.2多个总体的距离判别法

5.3贝叶斯（Bayes）判别法
5.3.1基本思想
5.3.2多元正态总体的Bayes判别法
5.4费舍（Fisher）判别法
5.4.1两组判别分析
5.4.2多组别费舍判别法
5.5逐步判别法
5.5.1引入和剔除变量所用的检验统计量
5.5.2逐步判别的原则
第六章　主成分分析
6.1主成分分析的基本原理
6.2主成分分析的推导
6.2.1从协方差出发求解总体主成分
6.2.2从相关阵出发求解总体主成分
6.2.3样本的主成分
第七章　因子分析
7.1因子分析的基本理论
7.1.1因子分析的数学模型
7.1.2因子模型中的几个统计特征
7.2因子载荷阵的估计方法
7.3因子旋转
7.4因子得分
7.5因子分析的步骤与逻辑框图
7.5.1因子分析的步骤
7.5.2因子分析的逻辑框图
第八章　典型相关分析
8.1典型相关分析的数学描述
8.2总体典型相关
8.3样本典型相关
8.4典型相关系数的显著性检验
8.5典型相关系数的步骤及实例
第九章　对应分析
9.1对应分析及基本思想
9.1.1对应分析的数据类型
9.1.2对应分析的基本思想
9.2列联表及列联表分析简介
9.3对应分析的基本理论
9.3.1距离与总惯量
9.3.2 R型与Q型因子分析的对等关系
9.4对应分析的步骤

第十章　多维标度分析
10.1距离阵和经典解
10.1.1欧式距离阵
10.1.2欧式距离阵的判定定理
10.1.3多维标度的经典解
10.2实例
第二篇　多元统计分析实验
实验一　均值向量和协方差阵的检验
1.1实验背景
1.2实验步骤和结果分析
实验二　聚类分析
2.1实验背景
2.2实验步骤和结果分析
实验三　判别分析
3.1实验背景
3.2实验步骤和结果分析
实验四　主成分分析
4.1实验背景
4.2实验步骤和结果分析
实验五　因子分析
5.1实验背景
5.2实验步骤和结果分析
实验六　典型相关分析
6.1实验背景
6.2实验步骤和结果分析
实验七　对应分析
7.1实验背景
7.2实验步骤和结果分析
实验八　多维标度分析
8.1实验背景
8.2实验步骤和结果分析
参考文献